Interactive Intent Modeling for Exploratory Search

Ruotsalo, Tuukka

2018-10

Ruotsalo , T , Peltonen , J , Eugster , M J A , Glowacka , D , Floréen , P , Myllymäki , P , Jacucci , G & Kaski , S 2018 , ' Interactive Intent Modeling for Exploratory Search ' , ACM Transactions on Information Systems , vol. 36 , no. 4 , pp. 44:1-46 . https://doi.org/10.1145/3231593

http://hdl.handle.net/10138/308883
https://doi.org/10.1145/3231593

acceptedVersion

Interactive Intent Modeling for Exploratory Search

TUUKKA RUOTSALO, University of Helsinki[†], Finland
JAAKKO PELTONEN, University of Tampere,[†], Finland
MANUEL J.A. EUGSTER, Aalto University, Finland
DOROTA GŁOWACKA, University of Helsinki, Finland
PATRIK FLORÉEN, University of Helsinki, Finland
PETRI MYLLYMÄKI, University of Helsinki, Finland
GIULIO JACUCCI, University of Helsinki, Finland
SAMUEL KASKI, Aalto University, Finland

Exploratory search requires the system to assist the user in comprehending the information space, and expressing evolving search intents for iterative exploration and retrieval of information. We introduce interactive intent modeling, a technique that models a user's evolving search intents and visualizes them as keywords for interaction. The user can provide feedback on the keywords, from which the system learns and visualizes an improved intent estimates and retrieves information. We report experiments comparing variants of a system implementing interactive intent modeling to a control system. Data comprising of search logs, interaction logs, essay answers, and questionnaires indicate significant improvements in task performance, information retrieval performance over the session, information comprehension performance and user experience. The improvements in retrieval effectiveness can be attributed to the intent modeling and the effect on users' task performance, breadth of information comprehension, and user experience are shown to be dependent on a richer visualization. Our results demonstrate the utility of combining interactive modeling of search intentions with interactive visualization of the models that can benefit both directing the exploratory search process and making sense of the information space. Our findings can help design personalized systems that support exploratory information seeking and discovery of novel information.

CCS Concepts: • **Information systems → Users and interactive retrieval**;

Additional Key Words and Phrases: Proactive search, user intent modeling

ACM Reference Format:
Tuukka Ruotsalo, Jaakko Peltonen, Manuel J.A. Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive Intent Modeling for Exploratory Search. *ACM Transactions on Information Systems* 0, 0, Article 0 (2018), 45 pages. https://doi.org/10.1145/3231593

[†]TR and JP contributed equally.

Authors' addresses: Tuukka Ruotsalo, University of Helsinki[†], Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Jaakko Peltonen, University of Tampere,[†], Faculty of Natural Sciences, Finland; Manuel J.A. Eugster, Aalto University, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Dorota Głowacka, University of Helsinki, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Patrik Floréen, University of Helsinki, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Petri Myllymäki, University of Helsinki, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Giulio Jacucci, University of Helsinki, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland; Samuel Kaski, Aalto University, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Finland.

1046-8188/2018/0-ART0
https://doi.org/10.1145/3231593

1 INTRODUCTION

Information retrieval research has primarily focused on improving retrieval for a single query at a time or a short sequence of queries occurring in a search session. However, many complex tasks, such as literature surveys or product comparisons, require the user to become accustomed to a wider body of knowledge to complete the task. The objective of an exploratory search is not only to retrieve relevant results for a particular query, but to aid the user in completing more complex tasks that may involve evolving and changing search intents. Therefore, the goal of an exploratory search system is not only to find a diverse set of results or to minimize the time the user has to spend to look for a small highly-relevant or obvious subset of results. A search system tailored for exploratory search should rather provide the user with adequate interaction affordances to allow exploration to cumulatively gain information during a search session.

Consequently, user behavior has been shown to depend on the type of task. For example, researchers have proposed a distinction of "lookup", "learning", and "investigation" task types [21, 75, 118]. Lookup tasks aim at finding information of which the user is already aware, and they are already well supported by the current generation of search engines. Conversely, systems supporting exploratory search, requiring learning and investigating the information space to comprehend a wider body of knowledge, have turned out to be more difficult to design. One reason is that in an exploratory search setting the user is not *a priori* familiar with the information space.

A user engaged in an exploratory search task is facing the relevance paradox and an anomalous state of knowledge [17, 18, 34]. This means that the relevance of a particular piece of information for the task may become apparent to the user only after it has been retrieved. This implies that the user has to know what to search for, as existing approaches do not support becoming aware of information needs that arise during exploration. Consequently, exploratory search processes require iteration between discovering and subsequently comprehending retrieved information throughout the course of the task [32, 48]. Empirical evidence also suggests that in a large portion of web search sessions users are struggling, and these sessions can be mostly characterized as exploratory search sessions [81].

The needs of the user may evolve throughout the search session, and the user may need assistance in directing the search to explore initially unpredictable but highly relevant information. In the current typed-query based search user interfaces, users are forced to invest significant cognitive efforts in acquiring cues to formulate queries from the intermediate results, instead of being able to focus on discovering, learning, and collecting relevant information [114].

Intent modeling can help to sequentially mitigate the relevance paradox, the initially anomalous state of knowledge, and consequently struggling to find information by acquiring knowledge of the task from user interactions over the duration of the task [73, 118]. Consequently, users can discover and explore information relevant to their tasks rather than optimizing results to be maximally relevant for an individual query, which may be suboptimal in the first place. Recent results suggest that systems leveraging knowledge about a user's search intentions can provide higher-quality task outcomes on a longer term, than systems that try to optimize the immediate results set against the present query [1, 41, 119]. For example, a scientist conducting a literature survey on an unfamiliar topic would need to learn new conceptualizations that are potentially relevant to the research topic, use them to explore related but still relevant information, and comprehend the displayed information to make sense of relevant and useful information for completing the task. A search system best supporting the scientist to accomplish such a task would allow a dialogue between the system and the user to not only retrieve relevant results for a particular query but to also assist in discovering related information, conceptualizing the relevant information space for comprehending the available information, and helping to specify search intents that evolve throughout the course of the task when users learn and gain information.

Despite the fact that a significant portion of search activities is associated with exploratory tasks, and task complexity is known to affect search success [21], current interaction methods for supporting exploratory search

are based on techniques that suggest terms or rephrased queries [62], document relevance feedback mechanisms [61], and faceted search interfaces that enable narrowing down the search within the initial query scope [104, 125]. These techniques put the user in a reactive role to filter search results rather than offering interaction affordances to actively direct the exploratory search process.

We introduce *interactive intent modeling* to support exploratory search [as discussed earlier in preliminary conference papers and an overview paper: 43, 82, 96–98]. It is based on three principles:

(1) **Intent modeling.** The intent model predicts the user's evolving information need during a search session in an interactive modeling-based loop. The exploration-exploitation trade-off of reinforcement learning is used to control exploiting the best estimates and explore alternative, yet relevant, directions in the information space and present them as suggestions for subsequent interactions.
(2) **Intent-model visualization.** The intent model is transparently visualized for interaction. The visualization of the model can turn the human memory recall task to a fluid visual recognition task. Instead of recalling queries from human memory, the user can visually recognize potential intents expressed as keywords and interact with them. The user can direct the search by reinforcing or penalizing selected keywords, even when the space of possible alternative intents is large.
(3) **Intent-model based retrieval.** The intent model provides a relevance weighting for the document features. This is used in retrieving an updated set of documents reflecting the change in the user's search intent.

The objective of the article is to investigate interactive intent modeling utilizing reinforcement learning with feedback from the visualization of the information space. We validate the approach through comparative user studies in exploratory search and information comprehension.

We report two experiments using a system implementing the approach on a database indexing over 50 million scientific documents. The first experiment focuses on the effects of interactive intent modeling on the user's task performance (measured by expert ratings of users' answers), the system's retrieval performance (measured by precision, recall, and F1, cumulatively over the course of the session and at the end of the session), and user interactions with the intent model in exploratory search tasks. The experiment shows improved system retrieval performance with all conditions that implement interactive intent modeling, but task performance is only improved when rich visualization is used. The second experiment focuses on the visualization and studies the information comprehension support of the visualization operationalized as a user's ability to find semantic topics that cover the information content in a set of search results. The experiment shows improved coverage attributed to the visualization of the model.

2 RELATED WORK

To investigate how to support information retrieval in complex and exploratory tasks, we start by reviewing related work on information seeking and exploratory search that originate from empirical observations of human search behavior. After that we relate our approach to more recent studies that describe opportunities and challenges in adapting search by mining and modeling interaction data acquired from search engine logs. Moreover, we highlight our contribution to recent work in the area of search intent modeling and personalization of search. We then turn to reviewing search user interfaces, in particular, interfaces that make use of information visualization to assist comprehending search results, and adapting and personalizing search.

2.1 Empirical and Theoretical Frameworks of Exploratory Search

Possibly the best known hypothesis explicating the roots of the exploratory search problem is the anomalous state of knowledge hypothesis [17, 18]. The anomalous state of knowledge hypothesis states that in many cases, users of search systems are unable to precisely formulate what they need as they miss some vital knowledge to

formulate queries. In such cases the system should attempt to model a user's intentions to assist the user rather than to require the user to specify a query explicating the information need [78].

As a result, exploratory search has been described as a combination of exploratory browsing with focused searching [16, 75], where mixed search strategies are used to achieve task goals and the user is explicating the information need in different phases. It has been shown that tasks, goals, users' pre-knowledge and task phases are factors in this process [32, 73].

Focused and exploratory searching strategies are also analogous to orienteering and teleporting strategies [114]. The orienteering strategy refers to search behavior where the user issues a quick, imprecise query, reaching approximately to the right region of the information space. Users then follow paths that require small steps that move them closer to their goal. Conversely, teleporting refers to a strategy where the user issues a more precise query to jump directly to the target. This requires more effort in query formulation and prior knowledge about the problem domain, as users have to be able to formulate the query more precisely already in the beginning of their search. For a long time, most of the search engines have focused on supporting teleporting. Behavioral studies, however, have demonstrated that a large portion, around 40% to 65% [114], of the search goals are informational, in which users want to learn about something they are not *a priori* familiar with. Yet, recent empirical evidence also suggests that in a large portion of web search sessions users are struggling, and these sessions can be mostly characterized as exploratory search sessions [81]. These behavioral findings highlight the importance of techniques and user interaction support that can assist users in orienteering to their goals and adjusting expressions of their information needs that arise during exploration.

Some classic frameworks have a more holistic view of information seeking. The information search process model describes users' experience in the process of information seeking as a series of affective, cognitive, and physical functions. Affective states that begin as uncertain, vague, and ambiguous become clearer, more focused, and specific as the search process progresses [68]. The early stages of the process model identify the uncertainty related to the information needs and the exploration phase in which inconsistent and incompatible information is encountered and new information discovered. Another framework directly related to our work is the berry picking model which refers to understanding search behavior as a multistage process of recognizing the problem, establishing a plan for the search, conducting the search, evaluating the results, and iterating through the process [15]. The information foraging theory has also inspired the development of exploratory retrieval techniques [83]. It draws an analogy to humans' evolutionary pressure to optimize their actions. This theory proposes that humans constantly make decisions on what kind of information to look for, whether to continue using the current source of information to try to find related additional information, or to move on to another direction, and when to finally stop the search. Although human cognition is not a result of evolutionary pressure to improve search, the analogy can be seen in users' aim to reduce cognitive effort to optimize search behavior [84]. These frameworks are partially overlapping and emphasize different aspects of humans' roles in approaching the information seeking problem. For example, a difference between exploratory search and information foraging is that in the latter hypothesis, users optimize their cognitive efforts, while in the former hypothesis, the users' willingness to invest time and effort to explore may change depending on what information is found during exploration [123]. While these models explain different aspects of the information seeking process, they all define some common aspects of the process: the initial anomaly in user's knowledge, the different phases of the search processes where users are trying to find, learn, and further specify their information needs and intentions, and the focus on information seeking episodes that can be described as a series of interaction between the user and the information that are supported by the information retrieval system.

2.2 Complex Tasks and Search Personalization

The majority of previous studies on practical information retrieval focuses on optimizing and evaluating look-up retrieval, where the information target is well-defined and human-machine interaction is limited to queries and search-result selections [90]. While look-up retrieval already serves many of our information needs, many practical tasks that generate information needs require interactive support to assist the search process [14, 75].

More recent work has highlighted the importance of interaction support for more complex tasks [118], whole-session relevance [90], and task performance beyond session boundaries [71]. Several studies have also shown how to predict search intents and intentional task types from query data and search behavior. Researchers have used data from user interactions, such as query history or interactions with a visual interface [9, 28, 45], search logs [12, 79], and search-interface dependent search-behavior features [78]. Studies show that both behavior inside one session and historic behavior over sessions can be used in combination or isolation to improve search results [19]. Also statistics about results repetition within search sessions have been incorporated into ranking for personalizing search results [106]. Further studies investigated how typical or atypical queries can be identified and help profile a search session to improve search results [40]. Other studies explored how to detect if queries have little ambiguity in intent but seek content covering a variety of aspects [90] as well as learning semantic query annotations suitable to the target intent of each individual query [42]. Such studies mark a trend towards personalized search based on models representing users' individual needs and intentions [77] that can model topical and even cognitive aspects of user intentions [60].

Users often struggle in formulating their intents as queries when they are engaged with exploratory search tasks. As a response, personalization techniques have been developed to support query formulation and relevance feedback. This is grounded in a well-known cognitive science theory stating that humans find recognition easier than recall [10], as it is usually easier for a human to recognize something presented than it is for her to describe it without any support. Several techniques exist to either support query formulation or to process results in order to help re-rank, filter [116], expand [31, 77], diversify [2, 27, 54, 89], or extract entity-oriented search intents to improve query suggestion and recommendation [39, 91]. Relevance feedback [61] and query and term suggestions [62] have also been proposed to improve short-term navigational search, but provide limited support for exploratory search.

Term and document relevance feedback mechanisms have been found to improve retrieval in laboratory studies [61], but much of the evidence from user studies indicates that relevance feedback features are not used, or if they are, they are unlikely to result in retrieval improvements [50]. There are two main reasons for this. First, relevance feedback directly affecting the query often leads to a context trap, i.e., after a few iterations of feedback, users have specified their context so strictly that the system can no longer propose anything new and users are trapped within the present set of results. Second, the requirement to explicitly select relevant and irrelevant documents or terms, when the system's responses are not immediately satisfying for the user, can prevent the user from actively engaging with the feedback mechanisms.

An alternative approaches requiring less active user involvement are query suggestion diversification and answer set diversification approaches that recommend alternative interpretations of a given query upfront [38, 87].

Methods have been proposed for modeling search intents for diversification and improvements over conventional diversification methods have been achieved by clustering query refinements for intent detection [100] and using click-through data to intent-aware diversification [53] and diversification that targets to predict novel suggestions [102]. Researchers have also developed diversification methods that can make meaningful query suggestions context-aware by taking into account the search or session context [22, 23].

Other techniques use query clustering to similar intent classes or hierarchical models of intents [30, 54] and suggest a diverse set of queries using models that utilize a short-term context using the user's behaviour

within the current search session, such as the previous query, the documents examined, and the candidate query suggestions that the user has discarded [63], the page context that the user has browsed [29], or the whole search session [92].

Recent research proposed implicit detection of intentions using pre-search context by monitoring the documents visited prior to performing a search [67]. Similarly, most of the previous work has focused on identifying alternative queries that bear a strong similarity or relevance to the original query or query history [109]. The suggested queries may therefore be good alternatives to the initial query or predicting the next query, but are not necessarily helpful in exploring outside the initial query scope. In exploratory search, search intents may shift and new intents arise as the search progresses. This requires the user to be able to choose and switch among a number of potential and even partially contradictory intents, instead of just allowing the system to converge towards a single target.

More generally, research in session search has benefited from the introduction of the Session Track at TREC (see e.g. [24, 59]), but has also limited the scope to the interaction types recorded in the TREC sessions. TREC sessions are also relatively short compared to the sessions that we have recorded in our task-based experiments.

We approach the intent modeling by using reinforcement learning [72, 111], where the user rewards the model through interaction and the predictions of the model are used to retrieve information and visualized for interaction. Conventional reinforcement learning assumes user states and a planning process to reach states for maximizing reward. In the bandit setting, the user is directly rewarding the model through interactions with the model as opposite to computing policy over state transition space. In this setting, the user does not need to have a specific goal in mind at the outset of the search but can gradually provide evaluative feedback on the system responses to reward the model while exploring the information space.

In our approach, the multi-armed bandit algorithm is used to suggest keywords related to the ones the user has preferred in the past (exploitation), but also keywords that are related and have high uncertainty (exploration). Consequently, the user can reward a direction already in the search focus (exploit), or reward an alternative and uncertain direction (explore). In this context, if the user consistently gives reward on very similar information, then the system will converge on this type of information. Conversely, if the user rewards a varied type of information, then the system will keep on exploring until the user decides to converge on a specific search direction.

As opposite to conventional relevance feedback [94, 101], the bandit approach is not limited to predictions within the information already been retrieved; the predictions of the keywords representing the intent model do not have to come from the set that the user has already seen, but rather from a model that can represent the data collection – and in principle could use even data from other sources than the collection being searched. The relevance feedback is interpreted as providing reward to the model as opposite to a highest ranked document representations as in conventional relevance feedback.

Similar to our modeling approach, multi-armed bandits have been utilized to model user preferences by learning diverse rankings for a single query based on clicking behavior [56, 88] and learning rankings from pair-wise document comparisons derived from implicit feedback [52, 126]. Other similar techniques include POMDPs that have been recently proposed for re-ranking [128] and session search [74].

2.3 Visual Interfaces in Search

The early studies in search user interfaces already demonstrated that even simple interaction support, such as faceted search interfaces, can be effective for tasks where the search goal is well defined and the success is measured based on the system's response to short interaction sequences [51, 125].

The user interface designs of search systems make trade-offs in complexity and richness of the information being presented to the user. The reduction of interface complexity can happen in several dimensions: choice of

the amount of information or search control options to be shown, organization of this information into categories or some visualization that can reduce the complexity for the user and allow for more efficient and effective navigation.

The main user interface oriented approaches to present richer information and allow navigation include filtering by facets [125], result visualization and navigation through clusters [49], and visual search [3]. While these approaches provide means for navigation, they are purely based on analysis of the document data, and do not take into account that search intents for the same query can be very different and can be learned during the search session [125]. Modeling search intent allows reducing the complexity of the search interface by narrowing down available feedback options to those most likely to be relevant for the user's information need. In complex tasks, however, the user's information needs evolve throughout the course of the search and the user's ability to direct the search to solve the task at hand is critical [43].

Recent search systems employ visualization of the resulting information to enable faster relevance judgment and effective feedback [47, 58, 70, 76, 117]. A variety of visualization approaches of search results have been explored, including multiple linked lists, scatter plots, graphs and their combinations [70, 110]. These types of visual search systems are distinguished from familiar query composition based systems by their emphasis on rapid filtering to reduce result sets, progressive refinement of search parameters, continuous reformulation of goals, and visual scanning to identify results [3]. There is mixed evidence of the effectiveness of the proposed techniques, but research suggests that when the interaction and predictive mechanisms are adopted they typically lead to improved efficiency [58].

Current visual-search approaches attempt to better support exploration in different ways: supporting sense-making by incrementally and interactively exploring the network of data [25], showing how visualization supports user involvement in the recommendation by providing a rationale behind suggested items [121], and visualizing relations of different queries and result sets [5]. Research demonstrating how to support users to view and manipulate the user models are rare [6, 7] and are limited for information seeking, as they employ conventional static ranked lists with limited focus on exploration. More recent work attempts to combine personalization of search with visualization approaches indicating a renewed interest in adopting advanced visual interfaces for exploratory search [4, 5, 65]. Despite extensive research in visualization and interactive support for search result ordering, there is no conclusive evidence that these visualization approaches lead to improved retrieval performance or search result comprehension in the hands of users [50, 105]. Studies have mainly focused on reporting either behavioral findings on the exploration behavior [36, 69], or finding more diverse information or increased coverage of variables in the exploration process [124]. Moreover, existing approaches focus on visualizing different data dimensions as separate widgets [3, 36, 124], but not topical similarity and relevance simultaneously.

A line of recent work has focused on interactive support for making sense of search results and several interactive interfaces have recently been proposed. ExplorationWall [65] is an interface that allows incremental exploration and sense-making of large information spaces by visualizing documents and related entities as search streams. PivotPaths [36] is another recently proposed interface for exploring faceted information resources by visualizing facets as paths and supporting pivot operations as lightweight interaction techniques that trigger gradual transitions between the facets. Syed and Collins-Thompson [112, 113] introduced a retrieval algorithm designed to maximize educational utility of a search system as opposite to conventional relevance optimization.

Interactive interfaces that enable transparent control on user models have recently become popular [8, 66, 95, 96]. The idea behind these approaches is that, as opposite to visualizing results, the user model is visualized and the user can interactively provide feedback on the search intentions using the visualization. Similar visual controls for user modeling have also been proposed for recommender systems [13, 35, 122].

Effective presentation of search results is important in exploratory information search scenarios where a user tries to gain understanding of a topic in order to retrieve more specific or related information in subsequent

search iterations [75, 114]. A simple ranked list may be sufficient in simple look-up search scenarios where users focus on finding one or a few highly relevant documents. However, broader coverage of search results is needed especially when the aim of the information seeking is not to look up an individual relevant document but to gain an overall understanding of varied information across multiple relevant documents. Only part of the information in each document may be relevant, and information content over multiple documents may be interrelated; the user must then comprehend not only individual documents but a wider coverage of different aspects of relevant information spread across documents. Successfully comprehending information content across search results lets the user relate the result documents to each other, the query, and the underlying information need, and to exploit the information content appropriately in further processing of the found information [64].

While the existing work highlights the utility and applicability of visualization techniques for interactively browsing information collections in general, we lack understanding of the benefits of the visualizations. Visualizations can possibly improve users' effectiveness or efficiency in understanding or comprehending the information collections, or alternatively act as complex proxies for simpler interactions that the users perform as a part of their information exploration processes.

2.4 Contributions

Several contributions set interactive intent modeling apart from the previous research:

(1) We present the principle of interactive intent modeling that allows the user to interact with the intent model visualization in order to provide feedback on the model. We also demonstrate the technical realization of the approach by using reinforcement learning with rewards obtained from user interactions. This is in contrast to previous research in intent prediction and personalization that have mainly focused on analyzing query logs for suggesting queries, providing document-driven visualizations, or diversifying ranked document lists.
(2) We demonstrate implementations of the technique as parts of practical information retrieval systems. Unlike majority of previous studies proposing modeling users' search intentions, we study the effectiveness and efficiency of the approach in user studies with real-world system implementations indexing a real-life data collection of over 50 million scientific documents.
(3) We empirically validate the performance of interactive intent modeling in exploratory search and information comprehension tasks. We report significant improvements of task performance, information retrieval performance over the session, information comprehension performance and user experience over a conventional search system. This is in contrast to previous approaches that typically evaluate models against log files or artificial session benchmarks that can not capture user dynamics emerging from experiment-dependent user interactions.

3 INTERACTIVE INTENT MODELING

Interactive intent modeling is a technique that models the user's evolving search intents over a search session. The model learns intent estimates from user feedback and visualizes them as keywords for interaction as shown in Figure 1. Consequently, interactive intent modeling forms an interactive loop between the system and the user in order to refine the user's search intentions and direct the search process.

At each iteration, a set of keywords is suggested to the user based on the feedback obtained in previous iterations. Given the evolutionary nature of exploratory search, it is important to exploit the feedback elicited from the user, but also to balance it with exploration. Users must be able to focus on a specific subset of the documents (exploit), but at the same time to be able to broaden their search to more general, but still highly relevant, documents (explore). This learning procedure is called the exploration/exploitation tradeoff of reinforcement learning.

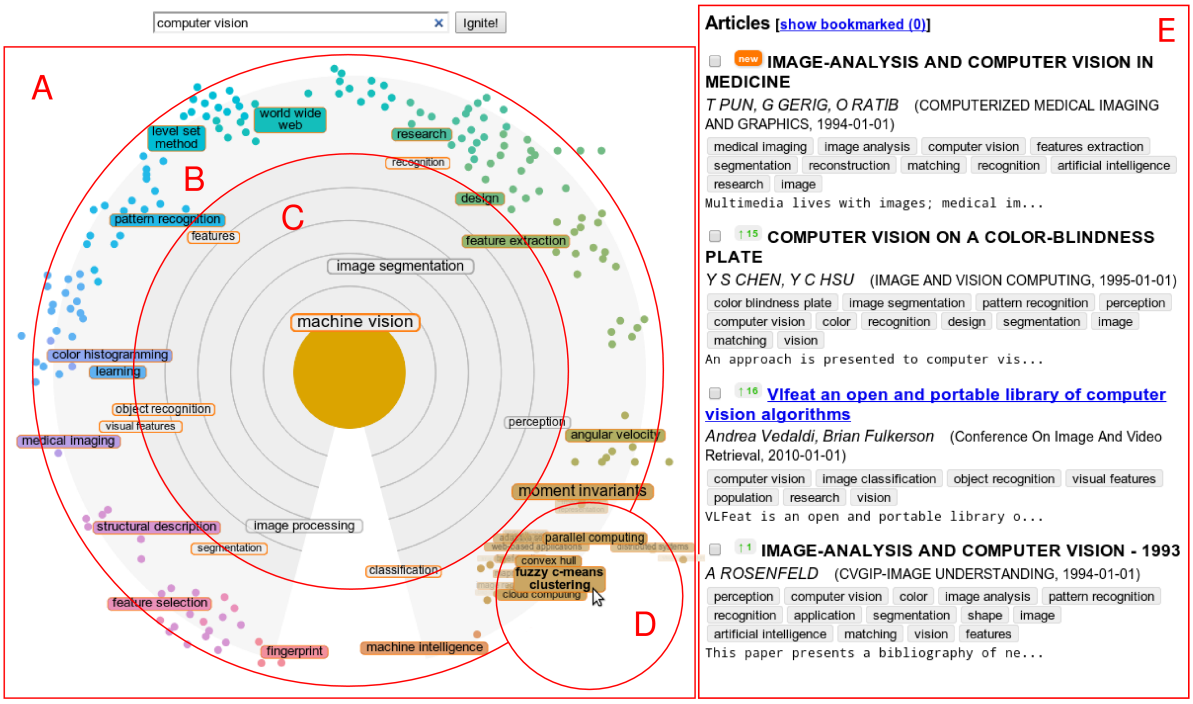

Fig. 1. A screenshot of the IntentRadar interface. A query, "computer vision", has been issued. Besides a query box and an article list, the interface also visualizes the predicted keywords representing search intent and potential future intents. The keywords are visualized with a set of keywords organized on a radial layout (A), where the center area represents the user: the closer a keyword is to the center the more relevant it is to the estimated intent. The present intent model used for retrieval is visualized as keywords in the inner circle (C), and the future projection of alternative search intents is visualized as potential new directions in the outer circle (B). Details of the visualization can be inspected with a fisheye lens (D). The ranked list of documents is shown on the right side of the interface (E).

Assisting the user in directing the search, the visualization shows both the current intent estimate, the alternative intents, and how the alternatives are related to the current intent estimate. The two-dimensional visualization shows the relevance of each keyword in the current estimated intent and the similarity of the keywords representing alternative intents.

3.1 A Walkthrough Example

We demonstrate an example implementation of interactive intent modeling in a system called IntentRadar, which indexes a large corpus of scientific documents. Interactive intent modeling represents the search intent through a set of weighted keywords associated with the documents, and the model is transparently visualized for interaction.

We explain the system and illustrate its usage via a walkthrough example of directing search in an exploratory search task. Figure 1 illustrates the IntentRadar interface when a query, "computer vision", has been issued. Besides a typical query box and article list, the IntentRadar interface provides an interactive visualization of the intent model on a radial layout. The center of the IntentRadar interface represents the user. In the inner area (**C** in Figure 1), keywords close to the center of the radar visualize the system's estimate of the user's present search intent. The outer area (**B** in Figure 1) consists of keywords that are not part of the present intent estimate but are

recommended for the user as potential future intents to explore. The interface thus allows the user to recognize potentially interesting intents to direct the search towards them.

The position of an individual keyword in the visualization is defined by the radius and the angle. In more detail, the *radius* of a keyword represents its relevance: the closer a keyword is to the center, the more relevant it is for the current estimated search intent. The *angles* of keywords represent their similarity: similar angles indicate the keywords are relevant to similar intents. To help distinguish topically different search intents from each other in the outer area, the interface also colors the keywords based on a clustering. The keywords with the highest relevance in each cluster are shown with labels to characterize the cluster. The other keywords are shown as dots that can be enlarged with a fisheye lens (**D** in Figure 1). The retrieved document list is visualized on the right side of the interface (**E** in Figure 1)

The visualization can be used to direct the search. Positive relevance feedback can be provided by dragging a keyword closer to the center of the radar, or by clicking a keyword under a document, which assigns full relevance for the keyword. Negative relevance feedback can be provided by dragging a keyword outside the radar. Feedback can be provided on several keywords at each iteration. After the user has finished providing feedback, the center of the radar is clicked by the user, and the system will update the intent model, the visualization, and the document list accordingly.

Figure 2 illustrates a sequence of exemplar interactions for a user who starts by issuing a query, "computer vision", and uses the interface to tune the search towards more detailed areas of interest. We focus on the user's interaction to direct the search. At any point during the interaction, the user can also read abstracts and visit the full articles.

After the initial query, the user is displayed a visualization (Figure 2, **A**) of the system's initial estimate of her search intent, potential future intents, and an initial set of documents. The user selects "object recognition" and "perception" to match her search intent and provides them positive relevance feedback by dragging those keywords towards the center of the radar (Figure 2, **B**).

The system learns from the feedback and updates the intent model to improve the estimates of the current and potential future intents. The visualization is correspondingly updated and a matching set of documents is retrieved (Figure 2, **C**). The user then further directs the search towards "stereo vision" and "tracking" by dragging those keywords closer to the center and receives an updated model and document list (Figure 2, **D**).

In the following we explain each component of the system presented in Figure 1. Section 3.2 explains estimation of the current intent model, which is used in part **C** of Figure 1. Section 3.3 explains the projections of future intents used in part **B** of Figure 1. Section 3.4 explains the layout computation for the visualization. Section 3.5 explains the document retrieval model that uses the intent model to produce a ranking of the documents listed in part **E** of Figure 1. Table 1 summarizes the notation used in the sections.

3.2 Estimation of the Intent Model

The purpose of intent model estimation is to predict a set of keywords and their relevances that together represent the user's search intent (part **C** of Figure 1) within a search session. Modeling intent over the search session has the advantage of capturing the user's overall search intent, while allowing the user to direct the search towards more specific topics.

The intent model estimation must balance exploration and exploitation of the user's feedback. If the system would simply estimate the intent by selecting the keywords with the highest exploitative estimates the system would suggest keywords similar to the ones presented to the user in the previous iterations. This would risk the user in getting stuck in a local "context bubble". Alternatively, the system could purely explore, which would lead to a set of "diverse" keywords, that is, keywords having large variance across the documents but having little to do with the user's feedback so far. Both situations are suboptimal given the evolving nature of exploratory search.

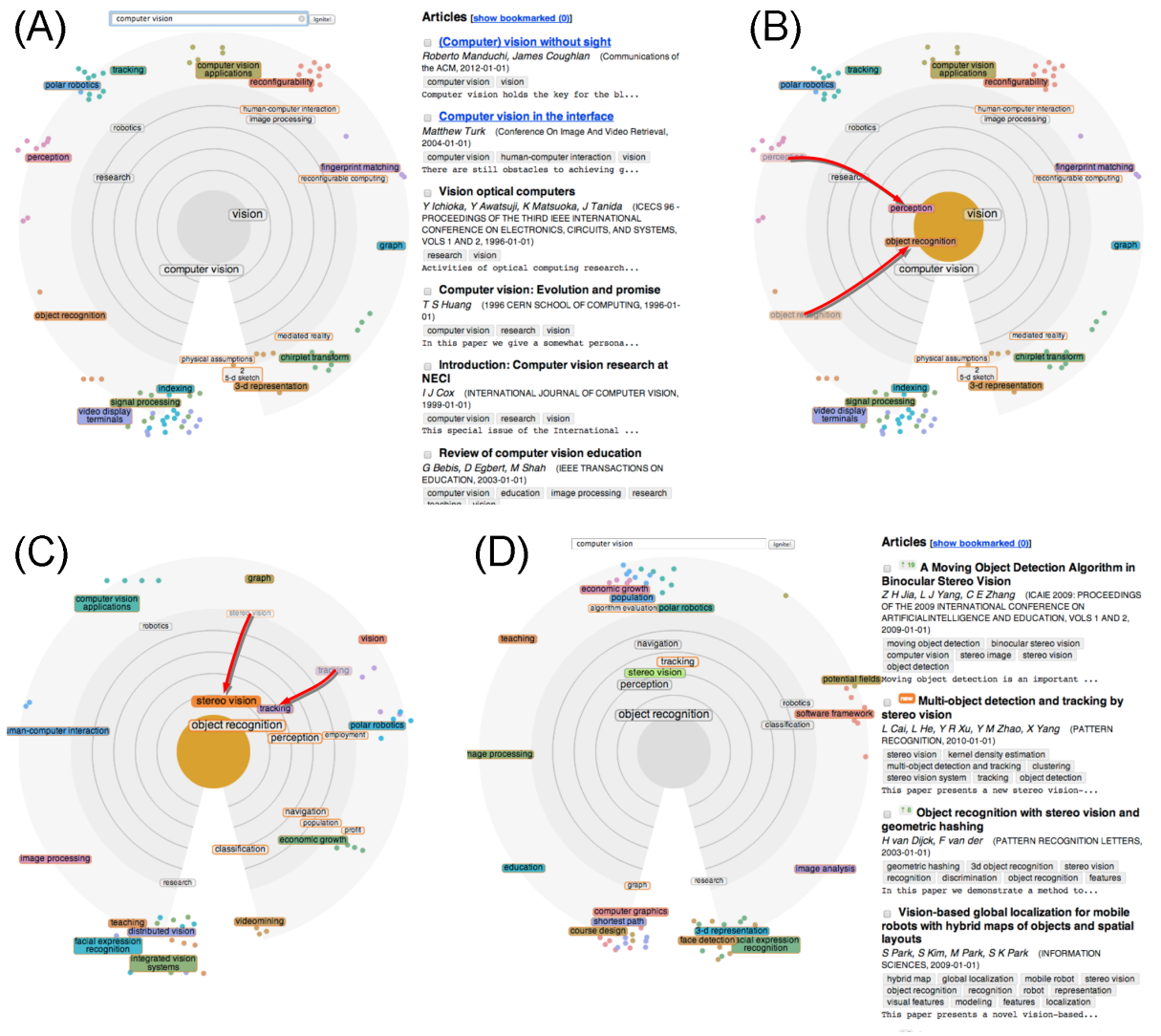

Fig. 2. An example of interactive intent modeling with the IntentRadar system. **A:** The initial visualization in response to a typed query "computer vision". The visualization shows keywords relevant to the estimated search intent and keywords for alternative future intents. **B:** The user wants to learn more about "object recognition" and "perception", and gives feedback by moving those keywords towards the center of the radar (movements are shown as small red arrows) and clicking the center of the radar. **C:** The user receives a new estimate of the present and future intents and an updated document set (omitted in the figure for brevity). The user wants to learn about "stereo vision" and "tracking" and gives feedback by moving the keywords closer to the center of the radar and clicking the center of the radar. **D:** The user again receives an updated intent estimate and an updated document list.

Symbol	Meaning
T	total number of iterations in the search session
t	index of an iteration, omitted where obvious
F	number of keywords that have received relevance feedback so far
r	relevance feedback value for a particular keyword, a number between 0 and 1
$\tilde{r}$	vector of all F relevance feedback values received so far
K	data matrix for keywords that have received feedback. Rows represent keywords and columns represent documents
k	vector for a keyword whose relevance is to be predicted
λ	regularization parameter in regression model
a	regression weight vector for a keyword whose relevance is to be predicted
$\hat{v}$	relevance score for a keyword: upper confidence bound of predicted relevance for the keyword
c	constant used to adjust the confidence bound to balance exploration and exploitation
L	number of alternative pseudo feedback considered
l	index of an alternative pseudo feedback
$\hat{v}^{\text{fu},l}$	upper confidence bound of predicted relevance for a keyword after the lth alternative pseudo feedback
i and j	indices of two keywords, omitted where obvious
$\tilde{v}_i$	vector of L predicted relevances for keyword i, corresponding to the L pseudo feedback; characterizes behavior of the keyword in response to feedback
$p(j\|i)$	similarity of keyword j to i as a probability, based on similarity of their characterizations
$q(j\|i)$	on-screen apparent similarity of keyword j to i as a probability, based on similarity of their angles
σ_i	parameter controlling falloff of the similarity probabilities around keyword i
α_i	angle of keyword i on the IntentRadar
p_i	probability distribution over keywords j based on the similarity of their characterizations to i; contains all values $p(j\|i)$
q_i	probability distribution over keywords j based on the similarity of their angles to i; contains all values $q(j\|i)$
D_{KL}	Kullback-Leibler divergence between two probability distributions

Table 1. Mathematical notation for intent model estimation, intent model visualization, and document retrieval.

In order to solve the exploration/exploitation tradeoff, we utilize the *LinRel* algorithm [11]. First, LinRel estimates a linear model representing the current search intent based on the search session history. Second, it predicts expected keyword relevances and corresponding *upper confidence bounds*. Third, keywords with high upper confidence bounds are selected for visualization. Intuitively, keywords with high upper confidence bounds

are the ones that are either already highly relevant with less uncertainty, or potentially relevant but with greater uncertainty. These are the keywords that are optimal for user feedback in order to improve the intent model.

In detail, a search session consists of $t = 1, \ldots, T$ iterations. At each iteration t, a large set of keywords have been assigned estimated relevance scores from the intent model; the top-k ranked keywords are visualized, and the user provides relevance feedback for one or more of the visualized keywords. The intent model is then re-estimated taking the new feedback into account, yielding new estimated relevance scores for all keywords.

The new estimated relevances are then used for retrieving documents (Section 3.5) and updating the visualization (Section 3.4) for the iteration $t + 1$.

Suppose up to time t we have collected F instances of relevance feedback where each feedback $r_1, \ldots, r_F \in [0, 1]$, is a relevance value for a particular keyword, and the vector of all feedback values is denoted $\tilde{r}$. The model operates on a *tf-idf*-valued [57] data matrix K where documents are columns and the F keywords that received feedback are rows.

The algorithm then consists of two steps. For any keyword (whether it has received feedback or not), denote by k the *tf-idf* vector of documents versus that keyword. For this keyword, the algorithm first computes a regression weight vector

$$a = k(K^\top K + \lambda I)^{-1} K^\top, \tag{1}$$

where I is the identity matrix, and λ is a regularization parameter set to 0.5. In the second step, the final relevance score at the current iteration is computed for the keyword, by taking into account the vector $\tilde{r}$ of feedback obtained so far:

$$\hat{v} = a \cdot \tilde{r} + \frac{c}{2}\|a\|, \tag{2}$$

where $\tilde{r}$ is the vector of feedback obtained so far, a is the regression weight vector for the keyword whose relevance score we are predicting, $\|a\|$ is the L_2 norm of the regression weight vector, and the constant c is used to adjust the exploration / exploitation trade-off (we used $c = 2$ to give equal weight for exploration and exploitation). This procedure is repeated for all keywords whose relevance needs to be estimated; note that in Eq. (1) all terms except k on the right-hand-side are the same for all keywords, thus the computation can be quickly done over all keywords whose relevances need to be estimated. It can be shown that this procedure is equivalent to estimating the upper confidence bound in a linear regression problem [11].

Intuitively, the first term on the right side of Equation 2 models exploitation and the second term exploration. Therefore, at each iteration of the search, based on the feedback, the model suggests not only keywords with the highest relevance score, but the keywords with the highest confidence bound, balancing exploration and exploitation as desired.

3.3 Projecting Future Intents

The keywords together with their corresponding relevances form the user's present intent model which is visualized in the inner part of the visualization (**C** of Figure 1). In order to offer the user these feedback options and to allow directing the search towards alternative, yet relevant intents, the system estimates how the relevances of keywords would change in response to simulated feedback corresponding to future intents. Keywords that may become relevant in the future intent predictions are then offered to the user in the outer part (**B** of Figure 1) of the visualization.

To predict a diverse set of future intents, we use several alternative pseudo feedback, one for each of the top L most relevant keywords; that is, if relevances $\hat{v}_1, \ldots, \hat{v}_M$ have been computed for a large set of M keywords, the keywords having the L largest values are each separately selected for pseudo feedback. In practice, a pseudo-relevance feedback with value 1 (strong positive feedback) is given as feedback for the keyword at rank l to

simulate the lth potential future intent. Using this pseudo feedback in addition to the feedback received so far, a corresponding relevance estimate in this future intent is computed for all keywords, with the the intent estimation procedure introduced in the previous section. This is repeated for each of the L alternative pseudo feedback. Therefore, each keyword will receive L estimated future relevances $[\hat{v}^{\text{fu},1}, \ldots, \hat{v}^{\text{fu},L}]$ where each relevance arises from one of the alternative pseudo feedback.

The relevance estimates of all M keywords in all L future intents are collected into a matrix $\tilde{V}_t \in [0,1]^{M \times L}$. The estimates of future intents are derived from the model and enable projections to different directions in the information space that may become relevant for the user in subsequent iterations.

3.4 Visualization of the Intent-model

We now describe a computational method for the purpose of laying out the top L keywords in the inner circle of the visualization (**B** of Figure 1) and the keywords representing future intents in the outer circle of the visualization (**C** of Figure 1).

We use a radial layout that has a good tradeoff between the amount of shown information and comprehensibility. A simple list of keywords would only use one degree of freedom and would not show keyword relationships, whereas higher than two-dimensional visualizations could make interaction with the visualization more difficult [37]. The radial layout also has a natural reference point representing the user in the center.

We first start by describing the outer circle (**B**). In Section 3.3 we described how, at each iteration of the interactive search process, each keyword has been computed a high-dimensional representation: a vector of L future relevances $[\hat{v}^{\text{fu},1}, \ldots, \hat{v}^{\text{fu},L}]$ where each relevance is in response to a particular pseudo feedback. For a keyword i, denote this vector by $\tilde{v}_i$. The norm $||\tilde{v}_i||$ represents the overall relevance of the keyword over different possible feedback, and we use it as the radius of the keyword on the radial layout. We then normalize the vectors, and for the remainder of the discussion we simply refer to the normalized vector $\tilde{v}_i/||\tilde{v}_i||$ as $\tilde{v}_i$. For each keyword i, the (normalized) vector $\tilde{v}_i$ characterizes how the keyword behaves in response to feedback. Keywords that have similar characterizations behave similarly in response to feedback, and therefore represent a similar direction for directing the search. We aim to present such directions to the user as angles on the radial layout, and accomplish this by dimensionality reduction from the high-dimensional characterizations to low-dimensional angles [99].

The angles α are computed by a non-linear dimensionality reduction to one dimension. We use a state-of-the-art nonlinear dimensionality reduction approach that has outperformed others in recent comparisons [120]. The approach is based on probabilistic modeling, and aims to preserve keywords with similar characterizations by giving similar angles to similar keywords.

For the dimensionality reduction, we define similarities as follows. Two keywords are related if their relevance grows in response to the same feedback, and we assess this simply by comparing their characterizations. Suppose there are M keywords in total. For each keyword i, similarities to other keywords $j \neq i$, $j = 1, \ldots, M$ are set as probabilities

$$p(j|i) \propto \exp(-||\tilde{v}_i - \tilde{v}_j||^2/\sigma_i^2) \; . \tag{3}$$

For the visualization display, similarities are defined analogously based on the angles α_i of keywords:

$$q(j|i) \propto \exp(-(\alpha_i - \alpha_j)^2/\sigma_i^2) \; . \tag{4}$$

Both probabilities are normalized to sum to one over the j, and the σ_i are set as by Venna et al. [120]. Given these definitions of similarities, the dimensionality reduction algorithm finds the optimal angles for the keywords by minimizing the difference between the probability distributions $p_i = \{p(j|i)\}_{j \neq i, j=1,\ldots,M}$ and $q_i = \{q(j|i)\}_{j \neq i, j=1,\ldots,M}$. The total amount of difference, for all keywords i, is measured by a sum of Kullback-Leibler divergences D_{KL} between the distributions,

$$\sum_i D_{KL}(p_i, q_i) + D_{KL}(q_i, p_i) \; , \tag{5}$$

and the angles are found by minimizing the divergence with respect to the α_i using a gradient descent algorithm.

Next, we describe the inner circle (**C**). In order to ensure consistency between the inner and the outer circle, the angles of the L keywords displayed in the inner circle are placed according to the angles of the $M - L$ keywords displayed in the outer circle. Thus the keywords of future intents are shown close to the respective keywords of the current intent that they are most related to. The radius is chosen to be directly proportional to the relevance estimate of the keyword: the top L keywords are positioned in the inner circle (the closer to the center the more relevant) and the remaining keywords in the outer circle. Finally, the keywords are colored based on agglomerative clustering applied to the angles of keywords.

3.5 Intent-Model Based Retrieval

The purpose of the retrieval model is to provide at each iteration t a list of ranked documents (**E** of Figure 1) given the estimated relevance scores $\hat{v}_{t,i}$ for the L most relevant keywords, i.e., the top-ranked keywords, which are also displayed for the user in the inner circle of the visualization (**C** of Figure 1) .

For each of the documents in the collection, we compute the relevance score, for which we use the language modeling approach of information retrieval. This approach computes the probability of the intent model generated by the given document [85]. To avoid zero probabilities and improve the estimation we use an estimate smoothed by Bayesian Dirichlet smoothing [127].

The final list of k documents is shown to the user and selected from the ranked list. The simplest solution to establish the final list of documents shown for the user would be to select the top-k ranked documents according to the language model ranking. However, to favor diversity, and to ensure the final results list represents the different intents present in the intent model, we diversify the ranked list of documents. The diversification is conducted by sampling a set of documents by using Dirichlet Sampling [44] from the top documents. To collect the session history, the new unique documents from the top ranked documents are added to K at each iteration of the search session.

4 EXPLORATORY SEARCH EXPERIMENT

The goal of exploratory search systems is to support users in discovering information to resolve an open-ended problem rather than maximizing short-term query-response performance. Therefore, a controlled task-based user experiment, that situates the participants in open-ended tasks, was designed.

4.1 Research Questions

The exploratory search experiment sought answers to the following research questions:

RQ1 Task performance: Does interactive intent modeling lead to better task outcome?
RQ2 Retrieval performance: Does interactive intent modeling result in high-quality retrieved information?
RQ3 Interaction support: Does interactive intent modeling elicit useful interactions?
RQ4 User experience: Does the increased complexity of the user interface design, compared to standard search interfaces, affect the subjective user experience?

4.2 Experimental Design

We chose a 3×2 between-subjects design with three system configurations and two tasks. This design was chosen to avoid the learning effects of participants, as each participant only used one of the systems and performed a single task with the system.

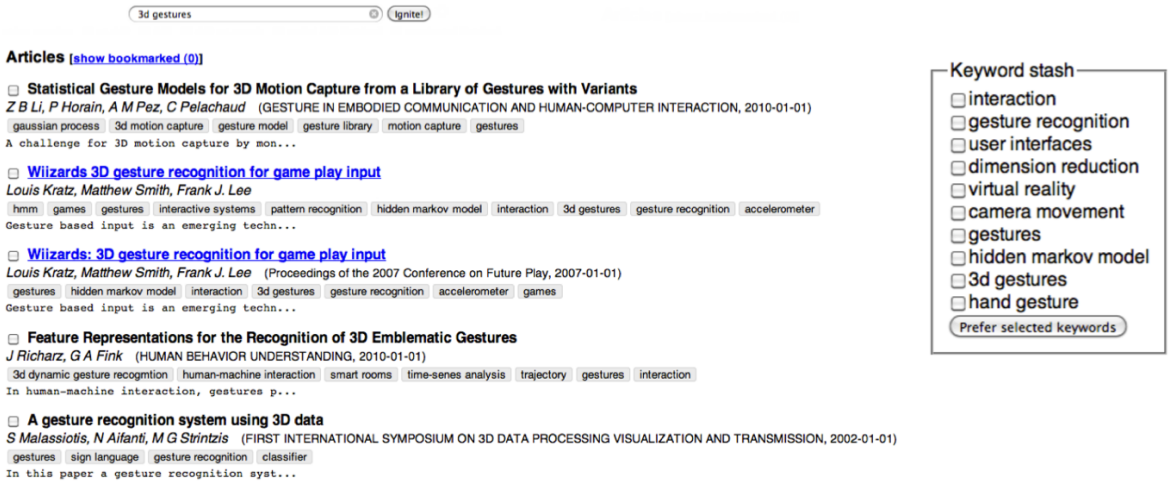

Fig. 3. A screenshot of the IntentList system. Relevance feedback can be provided by interacting with the estimated intents visualized as a list on the right side of the search result list.

4.3 System Configurations

Three systems were configured to study the roles of the intent modeling and the interactive visualization of the intent model. The system configurations and the associated features are as follows:

- **IntentRadar system** is a full system as presented in the previous sections and contained the intent modeling functionality and the radial visualization component.
- **IntentList system** is a system with the intent modeling functionality, but with a simpler list visualization of the top-10 keywords from the estimated intent model.
- **Typed Query system** is a baseline system, in which the interactions were constrained to typed-keyword queries and the results were presented as a ranked list.

A screenshot of the IntentRadar system is shown in Figures 1 and 2. A screenshot of the IntentList is shown in Figure 3, and a screenshot of the Typed Query baseline system in Figure 4.

The underlying document ranking model was the same for all three systems and all systems had the conventional typed-query interaction option: In all system configurations it was possible to type queries in the search box instead of using the intent model component.

4.4 Task and Topics

We chose a task type that is complex enough to ensure that exploration is necessary for participants to acquire the information to accomplish the task, and complex enough to allow participants to choose the kind of interaction that best supports solving the task. The task was expected to reveal exploratory search effectiveness both at the level of the system and in the interaction behavior.

The task was defined as a scientific writing scenario, i.e. the participants were asked to prepare materials and an outline for writing an essay on a given topic. The assignments were:

(1) Search for relevant articles to be used as references in the essay.
(2) Search for relevant keywords representing topics to be used to structure the essay.

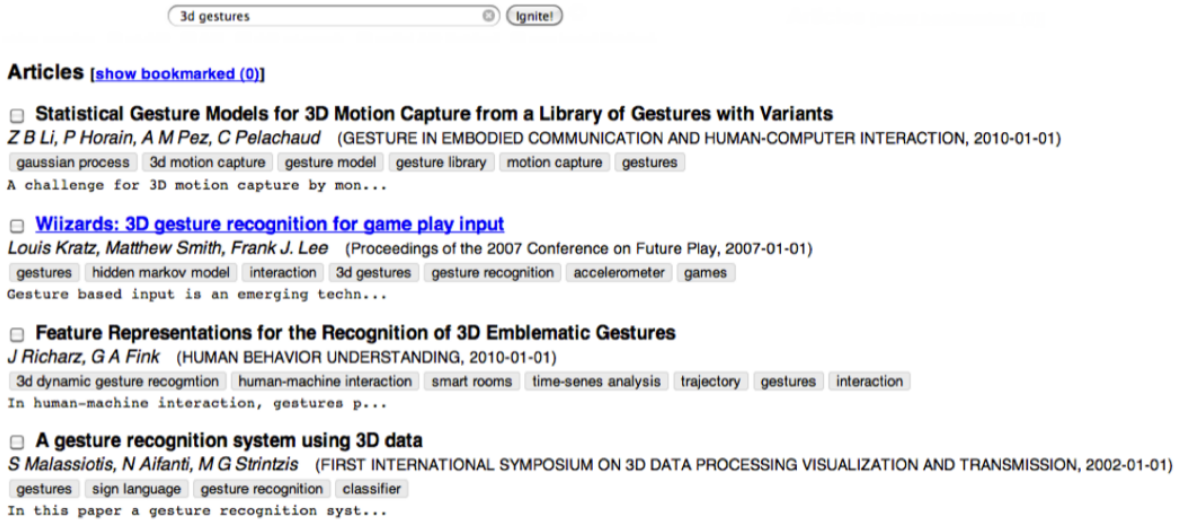

Fig. 4. A screenshot of the Typed Query baseline search user interface. Relevance feedback cannot be provided and the user is expected to reformulate the typed query to express alternative search intents. The Typed Query baseline system reflects standard search engine usage, where the interaction is via typing queries.

We recruited two post-doctoral researchers to define the topics and specific assignments for these topics. The topics chosen by the experts were "semantic search" and "robotics". The experts wrote task descriptions using the following template: "Imagine that you are writing a scientific essay on the topic. Search for scientific information that you find useful for this essay". In order to provide clear goals for exploration, the experts were asked to provide questions about specific aspects of the topic. The question defined by the experts for the robotics topic was: "What are the sub-fields, application areas and algorithms commonly used in the field of robotics?", while the question for the semantic search topic was: "What are the techniques used to acquire semantics, methods used in practical implementation, organization of results, and the role of semantic Web technologies in semantic search?". The participants were asked to both search for documents to support their answers to these questions and to write short answers under each question to fill in the essay outline.

4.5 Participants

We recruited 30 participants from two universities to participate in the study. The participants were 20-40 years old. There were nine female and 21 male participants. All the participants were graduate students with a technical background. Through a prior background survey we ensured that every participant was familiar with the concept of a literature search and had conducted one in their past experience.

We also screened the participants to avoid bias caused by pre-knowledge levels. The screening allowed to avoid recruiting participants who might be highly knowledgeable about a topic, e.g. an expert on robotics who would know almost all literature prior to the experiment, or an novice who might not know anything, e.g. a first year student without technical knowledge about the topic. The participants with high and low prior knowledge of the topic of the assigned search task were not allowed to participate. Prior knowledge was assessed via self-assessment on a scale of 1 to 5 ((1) no knowledge at all, (2) some knowledge, (3) moderate knowledge, (4) knowledgeable, (5) expert knowledge). We only allowed students to participate if they rated their prior knowledge between 2 and 4.

4.6 Procedure

The basic protocol for each experiment scenario was the following: instructions and demonstration of the system (10 min); processing of the search task by the participant (30 min); and completion of two questionnaires (10 min).

Prior to the experiment, the participants were asked to read written instructions. The instructions explained the task that the participants were expected to perform. The participants were then demonstrated the system. Then the participants watched a 2-minute video illustrating an exemplar task with the system variants and they had a 3-minute trial using the systems via a pre-defined query that was not related to the topics used in the actual experiment.

After this phase, the actual experiment started. In the experiment, the participants had full freedom to use the provided search system as they wished. The participants had 30 minutes to complete the task. The participants were notified 5 minutes before the end of the experiment to ensure that they were able to compete their essays before the end of the experiment. After the experiment, the participant filled in post-task questionnaires selected from the ResQue questionnaires [86].

4.7 Apparatus

The experiments were performed in an office-like environment using standard equipment (20"–24" monitor, mouse, and keyboard). The demonstration of the system was done by the instructor using a separate computer.

During an experiment all interactions by the participants with the systems were logged with timestamps, including typed queries, the documents and keywords presented by the system in response to interactions, and all interactions with the interactive components of the systems.

4.8 Data and Settings

We used a dataset of over 50 million scientific documents from the following data sources: the Web of Science prepared by THOMSON REUTERS, Inc., the digital library of the Association of Computing Machinery (ACM), the Digital Library of the Institute of Electrical and Electronics Engineers (IEEE), and the digital library of Springer. The dataset contains the following information about each document: title, abstract, keywords, author names, publication year and publication forum. Both systems used the same document set.

For the experiment we fixed the parameters in the system to the following values. The number of retrieved documents at each iteration by the language model ranked was set to 300 and we used $\mu = 2000$ for the Dirichlet smoothing. The number of documents shown to the user at each iteration was set to 20. The number of keywords included in the present intent model was set to $L = 10$. The maximum number of keywords representing future intents displayed for the user was set to 200.

4.9 Relevance Assessments

After the completion of the experiments, the experts who designed the tasks conducted two types of relevance assessments. First, assessments were done for the quality of information displayed and second, the quality of responses of the essay materials and outlines created by the participants.

All documents and keywords that were retrieved and displayed for the participants by any of the systems during the tasks were pooled, resulting in a pool of 5612 documents and 4097 keywords. Out of the 5612 documents, 3384 were labeled as relevant. Out of the relevant documents, 731 were labeled as obvious and 2653 were labeled as novel. Out of the 4097 keywords, 2225 were labeled as relevant. Out of the relevant keywords, 1284 were specific and 938 were general. Each document and keyword was assessed by two experts who created binary assessments of the documents on the following assessment categories:

(1) Relevance—is this article relevant to the search topic?
(2) Obviousness—is this a well-known overview article in the given research area?

(3) Novelty—is this article uncommon yet relevant to the given topic or specific subtopic in a given research area?

The experts also assessed the keywords by using the following assessment categories:

(1) Relevance—is this keyword relevant for the topic?
(2) General—does this keyword describe a relevant subfield?
(3) Specific—does this keyword describe a relevant specifier for the subfield?

The obvious and novel sets were disjoint and their union was the set of relevant documents. Analogously, the general and specific sets were disjoint and their union was the set of relevant keywords.

The quality of essay materials and outlines created by the participants were also assessed by experts. All the essay outlines from the participants were pooled and experts assessed each aspect of the essay, according to the questions specified in the task description, on a six point scale from 0 (no answer) to 5 (perfect answer).

A three-step process was used separately for documents and keywords. First, the assessors provided binary relevance assessment for each item. Second, all relevant items were categorized in the subcategories. Finally, all categorizations were checked again. All assessment procedures were double blind meaning that the assessors did not know the participants or the treatment conditions. To measure the inter-annotator agreement between the two experts, an overlapping randomly sampled subset of 10% of the articles and keywords was assessed by two experts. The Cohen Kappa test indicated a substantial agreement between the experts (Kappa = 0.71393, $p < 0.001$).

4.10 Evaluation Measures

Measures were defined to quantify each evaluation aspect defined in the research questions.

Task performance (RQ1) was the main measure of success. It was measured as the mean score of the expert grading of the participants' essay outlines.

Retrieval performance (RQ2) was measured by temporal and cumulative variants of the conventional effectiveness measures of information retrieval. We generalized precision, recall and F-measure to take into account the temporal dimension as new information was cumulatively gained while searching. Standard evaluation measures were not directly feasible because in our setting participants were shown a limited set of documents in response to an interaction, i.e., top-20, on each iteration instead of the total ordering. We also did not choose to use session-level discounted cumulative gain [55] as that measure penalizes the documents ranked lower and found later in the search session, which is against our intuition of evaluating the whole-session outcome. The goals of an exploratory search system are not minimizing time-on-task or the quality of an individual query-response, but information gain over the session. Consequently, we focus on measuring the gain of information within the search space and operationalize the evaluation as temporal recall and precision. Temporal recall and precision measures were adjusted to capture the performance of the system as a function of time. They measure the proportion of relevant documents cumulatively retrieved by the user up to a certain time point, in response to interactions with the system.

We start with a definition of cumulative document set $Pres_t$ denoting all unique documents presented to the user at a time point t, the set $Presrel_t$ denoting all unique relevant documents presented to the user at time point t, and $Allpres$ denoting all unique relevant documents in the assessment pool constructed from all documents found by any of the participants in the experiment. We define temporal precision as:

$$P_t = \frac{Presrel_t}{Pres_t}, \tag{6}$$

which measures the proportion of relevant documents cumulatively shown to the user, compared to all documents cumulatively shown to the user until time point t. Similarly, the the temporal recall is defined as:

$$R_t = \frac{Presrel_t}{Allpres}, \tag{7}$$

which measures the proportion of relevant documents cumulatively shown for the user compared to all relevant documents in the assessment pool.

For example, if a user finds relevant documents on the first iteration, just ten seconds after starting to use the system, but the system does not assist the user to direct the search, then the performance may not be much better when investigated after 120 seconds of use. On the other hand, if the system assists the user to gain more relevant documents, the recall after 120 seconds may have been increased because the user could easily interact with the system to gain more relevant results.

For example, by setting t = 60 seconds, we could investigate how many relevant articles a user was able to collect during the first minute of use by measuring R_t and what proportion of the collected articles were relevant by measuring P_t.

These measures are used to investigate how the precision/recall tradeoff develops over time and to compare systems in task settings where participants may use varying queries and interactions at varying points of time. The measures were computed with respect to the different assessment aspects: relevant, novel, and obvious.

Interaction support (RQ3) for directing exploration was measured using three separate types of measures. First, we measured the amount and type of interactions: typed query or interaction with the intent model. Second, we measured the quality of the keywords displayed for the user and the quality of keywords that the participants interacted with. Third, we measured the time as a function of the richness of the visualization, i.e., comparing the amount of keywords on the screen and the duration to interaction when this amount of keywords was presented. The hypothesis was to reveal whether the increased amount of information displayed would cause the participants to spend more time scanning to make the decision to which direction to explore.

User experience (RQ4) was measured using two standard post-test questionnaires: the standard System Usability Survey (SUS) [20] and ResQue, a recently proposed user-centric evaluation framework designed for the evaluation of recommender systems [86]. ResQue was chosen because it can be used to measure the subjective experience of interaction adequacy and preference expression capabilities offered by a system. The visualization system can be seen as a variation of a recommender system as it offers the users additional information to direct the search and present information. Participants filled in the questionnaires after completing the task and used a 5-point Likert scale to provide their answers (Strongly disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly agree (5)).

5 RESULTS OF THE EXPLORATORY SEARCH EXPERIMENT

The results show that interactive intent modeling, as implemented in both the IntentList and the IntentRadar systems, yields significantly improved retrieval performance and interaction support, and modest, but significant gain in user experience. The visualization and increased amount of information visualized in the IntentRadar system yielded a higher amount of interactions and improved task performance. These main findings are illustrated in Figure 5 and discussed from several points of view in the following sections: task performance, retrieval performance, interaction support, and user experience.

5.1 Task Performance

The participants who used the IntentRadar system achieved significantly better task performance than the participants who used the IntentList system or the Typed Query baseline system. The participants' responses to the tasks were graded significantly higher by experts. The mean expert grade for the IntentRadar system was

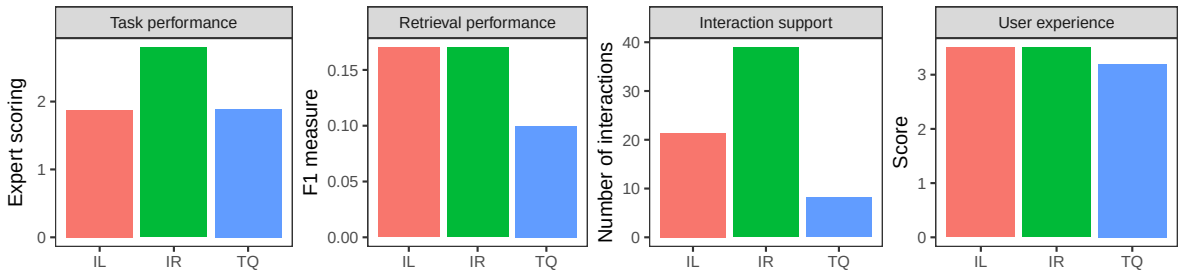

Fig. 5. Key performance measures for the mean performance of an average user at the end of the task: task performance, retrieval performance, interaction support, and user experience. All differences between the IntentRadar system and the Typed Query baseline are statistically significant, retrieval performance, interaction support and user experience differences between the IntentList and the Typed Query baseline are statistically significant, but in case of the task performance the the IntentList and the Typed Query systems perform equally. See the following sections for test details.

$\mu = 2.79$, for the IntentList system $\mu = 1.87$ and for the Typed Query baseline system $\mu = 1.88$. The differences between the IntentRadar and the other systems were statistically significant (Two-sided Wilcoxon rank sum test between the IntentRadar and the IntentList, $W = 1025.5, p = 0.007$, and between the IntentRadar and the Typed Query baseline systems, $W = 937, p = 0.005$). Both differences are significant using Bonferroni adjusted alpha levels of $p < 0.0167(0.05/3)$). Significant difference was not found between the IntentList and the Typed Query baseline system (Two-sided Wilcoxon rank sum test between the IntentList and the Typed Query baseline systems, $W = 691.5, p = 0.925$).

5.2 Retrieval Performance

Table 2 summarizes the retrieval performance in terms of the effectiveness measures for the compared systems for relevant, obvious, and novel categories. The values represent the effectiveness at the end of the task (at 30 minutes). Precision varied between 0.65 and 0.79 for the relevant category, between 0.26 and 0.35 for the obvious category and between 0.32 to 0.46 for the novel category. Significant differences were not found between the systems for precision of the retrieved documents in any category nor for any measure in the obvious category. Recall varied between 0.06 and 0.10 for the relevant category, between 0.13 and 0.16 for the obvious category, and between 0.03 and 0.09 for the novel category.

Both of the systems with interactive intent modeling show significantly higher F1-measure (IntentRadar $W = 10, p = 0.002$ and IntentList $W = 15, p = 0.007$) and recall (IntentRadar $W = 12, p = 0.005$ and IntentList $W = 16, p = 0.009$) for the novel documents than the Typed Query baseline system without sacrificing precision. These differences are significant using Bonferroni adjusted alpha levels of $p < 0.0167(0.05/3)$.

On average, participants found around 15% (109.65) of the obvious documents and between 3% to 9% of the novel documents. While the absolute recall may seem to be low, the measure counts for the overall effectiveness over the session as opposite to the conventional recall measure that counts for the mean individual query effectiveness. As a consequence, the actual count of the retrieved relevant documents has a significant variance between the systems. For example, in the case of novel documents the participants who used the Typed Query baseline system retrieved on average 79.6 (3% recall) novel documents, while the participants who used the IntentRadar system retrieved on average 238.7 (9% recall) novel documents.

These results indicate that while the participants were able to retrieve equally relevant and equally obvious documents, the participants' ability to explore to novel, yet relevant, areas was significantly improved when they were interacting with a system that employed interactive intent modeling. The effect sizes (F1-measure for

System	Precision			Recall			F_1		
	Rel	Obv	Nov	Rel	Obv	Nov	Rel	Obv	Nov
IntentRadar	0.65	0.26	0.40	0.10	0.15	**0.09**	0.17	0.17	**0.14**
IntentList	0.79	0.33	0.46	0.10	0.16	**0.08**	0.17	0.19	**0.13**
Typed Query baseline	0.71	0.35	0.32	0.06	0.13	0.03	0.10	0.18	0.05

Table 2. Retrieval Effectiveness for the Compared Systems in Terms of Precision, Recall, and F1 Measure and Computed for the Different Assessment Categories: Relevant, Obvious, Novel. Bold entries denote significant differences between the Typed Query baseline system and the compared system. See the main text for test details.

IntentRadar novel $\mu = 0.14, \sigma = 0.07$ and for IntentList $\mu = 0.13, \sigma = 0.08$) when compared to the Typed Query baseline system (F1-measure $\mu = 0.05, \sigma = 0.02$) are substantial and similar effect sizes hold for both systems that incorporate interactive intent modeling.

5.3 Temporal Performance Analysis

Retrieval effectiveness was measured after completing the task, i.e., measuring what the participant retrieved with the system during the 30-minute session. To gain more insights into the retrieval effectiveness, we analyzed the temporal effectiveness within the search session. Figure 6 (9 subfigures) shows temporal retrieval effectiveness of the compared systems with respect to precision, recall, and F_1 on the three ground truth aspects: relevant, obvious, and novel.

The results suggest that precision stays relatively constant for relevant documents throughout the session, but is slightly decreasing in the case of obvious documents. This is an intuitive result based on the experiment: since all users started the sessions by simply typing the name of the topic as the initiator query, the initial set of information contains many obvious documents, whereas after the initial set of obvious information had been retrieved, other interaction support becomes more important; as participants use such interactions to go beyond the initial obvious results, precision with respect the obvious documents will naturally decrease, whereas the fact that the overall precision of all relevant documents stays relatively constant means the system has allowed users to retrieve relevant documents beyond the obvious ones.

An interesting insight is that for the IntentRadar system, precision is slightly increasing towards the end of the session for novel documents. This suggests that richer interaction becomes crucial to discover novel information, in particular for these exploratory tasks that were studied in the experiment. The recall for relevant and novel information is increasing already after 250 seconds, suggesting that users can benefit from the intent modeling in much shorter sessions than our 30 minute test setup, but also suggests that interactive intent modeling has limited benefits in short look-up sessions. This is intuitive as if the users can successfully create a typed query for which the system can respond with few highly-ranked documents, the benefits from intent modeling and interactive visualization are limited. However, temporal analysis suggest that interactive intent modeling shows improvements for long-lasting exploratory search sessions, where users' goals are vague and evolving as they discover new information.

5.4 Interaction Support

Interaction support was first quantified by the amount of interaction to express information needs that was elicited by a system: typed queries or interactions with the intent model. Query reformulations were counted as queries, and interaction with the visualization were counted only when the dragging was successfully completed.

The left side of Figure 7 shows the total amount of interaction elicited by the different systems. The results show that participants adopt and make use of interactive intent modeling. The participants who used the IntentRadar

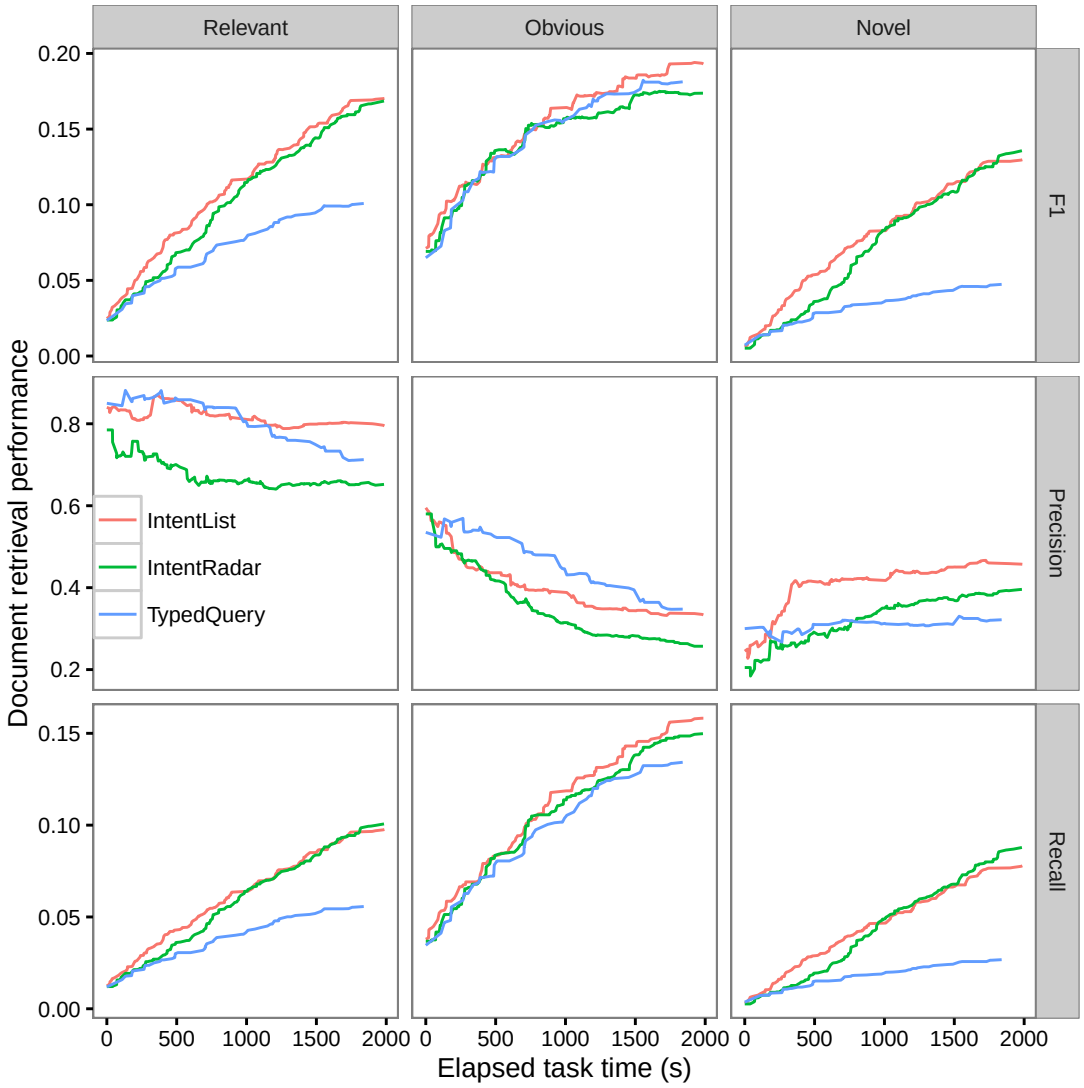

Fig. 6. Retrieval effectiveness of the compared systems in terms of precision, recall, and F-measure averaged over the participants and tasks, with respect to the elapsed task time. The three top figures show F_1 values for relevant, obvious, and novel documents, the three figures in the middle show precisions, and the bottom three figures recalls, respectively. No significant differences were found between the systems in regard to precisions, nor retrieval of obvious information. The systems with interactive intent modeling support achieved significantly improved effectiveness for novel information measured for the whole session duration.

system interacted four times more ($\mu = 38.9$) than the participants who used the Typed Query baseline system ($\mu = 8.7$) and the participants who used the IntentList system interacted nearly twice as much ($\mu = 21.2$) as the participants who used the Typed Query baseline system. The differences between the systems with interactive intent modeling were found to be statistically significant (two-sided Wilcoxon rank sum test, IntentRadar vs.

Typed Query baseline $W = 4223$, $p = 0.00005$, IntentList vs. Typed Query baseline $W = 3536.5$, $p = 0.0001$). Both differences are significant using Bonferroni adjusted alpha levels of $p < 0.00033(0.001/3)$). Despite the higher amount of interactions of the participants who used the IntentRadar system, no statistically significant difference was found between the IntentRadar and the IntentList systems. This suggests that the intent model improved the interaction of the users, but the different types of intent model visualizations did not affect the interaction behavior.

In order to study the distribution of different types of interactions that the different systems elicited, the interactions were split into the two types logged: typed queries and interactions with the intent model. The mean amount of typed queries was not found to be significantly different between the systems. The mean amount of typed queries for the IntentRadar was ($\mu = 7.8$), for the IntentList ($\mu = 7.1$) and for the Typed Query baseline ($\mu = 8.7$). The systems where the participants used the intent models for their interactions were used in cycles in which typed keyword queries were first issued and then interaction with the intent model was used to direct the search. However, the primary interaction method, for example, for the IntentRadar system was interaction with the intent model by interacting with the visualized keywords ($\mu = 31.2$) which is over three times more common than typed query interaction ($\mu = 7.8$). This indicates that participants did not replace the typed queries with interaction with the intent model, but rather directed their search further from the initially issued potentially imprecise query.

The middle panel of Figure 7 shows the amount of information on the screen as a function of time spent between interactions. Query typing and query reformulations were counted when the queries were issued. Interaction with the visualization were counted only when the dragging was completed. The dragging or typing time was excluded. In the Typed Query baseline system, the participants were always shown 20 documents and keywords associated to the documents were shown under each document. In the IntentRadar interface, the participants were shown 20 documents and keywords in the radar, and keywords associated to the documents were shown under each document. The right side of Figure 7 shows that the participants were able to provide the feedback in faster loops using the systems with interactive intent modeling despite having more information on screen. The mean duration between interactions for the participants in the IntentRadar condition was $\mu = 62$s, in the IntentList condition $\mu = 59$s, and for the Typed Query baseline system $\mu = 113$s. The difference between the IntentList and IntentRadar systems was not statistically significant, but the differences between the IntentList (Two-sided Wilcoxon rank sum test, $W = 3536.5$, $p = 0.0001$) and IntentRadar (Two-sided Wilcoxon rank sum test, $W = 4223$, $p = 0.000005$) systems compared to the Typed Query baseline system were significant. Both differences are significant using Bonferroni adjusted alpha levels of $p < 0.00033(0.001/3)$). This suggests that the participants were able to utilize the intent model rapidly and despite the increased amount of information on screen, the time to make decisions in directing the search was faster than in the Typed Query baseline condition.

As the amount of interaction was significantly increased and participants chose the IntentRadar as the main interaction method to direct their search, it was important to investigate the relevance of the keywords that were displayed and which the participants used in their interactions. Table 3 shows the precision, recall and F1-measure of the displayed keywords and the keywords that the participants interacted with. Participants who used the IntentRadar system achieved on average the recall of 0.3, which indicates that in the search session participants are able to cover approximately one third of the relevant keyword space. Participants who used the IntentList system covered only 0.06 of the relevant keywords. More importantly, the participants interacted with less and less of the relevant keywords using the IntentList system. In both systems, the participants interacted with relevant keywords as indicated by precisions of 0.84 and 0.96, and the precision of relevant and general keywords is higher for the keywords that were interacted with than it is for the displayed keywords. This suggests that even though the participants were shown more keywords in the IntentRadar system, they were able to select relevant keywords from the display. Notably, the higher recall of displayed keywords also did not cause

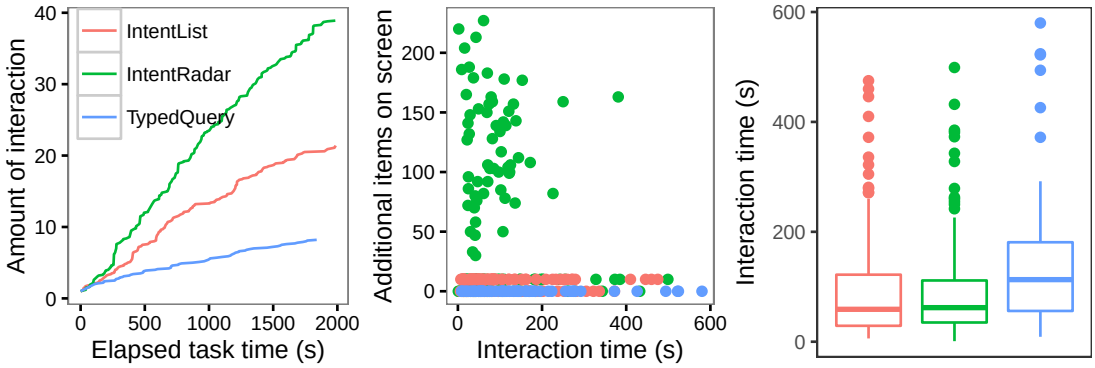

Fig. 7. Left: Amount of elicited interaction with respect to the elapsed task time. Middle, right: Time between interactions in relation to the number of documents and keywords displayed (middle) and summarized per system (right). The participants who used the IntentRadar interface had richer keyword visualization, interacted more and primarily with keywords, without affecting the time between interactions.

	Precision			Recall			F_1		
Cumulative keyword effectiveness (displayed):									
	Rel	Gen	Spe	Rel	Gen	Spe	Rel	Gen	Spe
IntentRadar	0.65	0.19	0.45	0.30	0.34	0.29	0.39	0.23	0.33
IntentList	0.80	0.29	0.51	0.06	0.08	0.05	0.11	0.13	0.09
Cumulative keyword effectiveness (interacted):									
	Rel	Gen	Spe	Rel	Gen	Spe	Rel	Gen	Spe
IntentRadar	0.84	0.43	0.41	0.02	0.01	0.01	0.04	0.02	0.02
IntentList	0.96	0.43	0.54	0.01	0.004	0.004	0.02	0.01	0.01

Table 3. The Keyword Effectiveness in Terms of Precision, Recall and F-measure of Displayed Keywords and the Keywords that the Participants Interacted with, for the Different Assessment categories: Relevant, General, Specific. The results are reported only for the IntentRadar and IntentList systems as the Typed Query baseline system did not enable keyword visualization or interaction.

longer interaction durations as shown in Figure 7, but participants were able to react equally fast to the presented information in all compared conditions.

5.5 User Experience

Analysis of the questionnaire data further supports the benefits of interactive intent modeling. General usability measured using the SUS questionnaire was perceived to be relatively high for all systems, but no significant differences could be found between the systems based on SUS questionnaires (SUS score for the Typed Query baseline was 65.7, SUS score for the IntentList system 65.7, and SUS score for the IntentRadar system was 60).

The ResQue questionnaires shown in Table 4 revealed significant differences between the systems. The mean user experience based on normalized answers of ResQue questionnaires (i.e., scores from the negative questions were inverted so that higher is always better) was found to be significantly higher (two-sided Wilcoxon rank sum test with Bonferroni corrected Alpha, $W = 347$, $p < 0.05$) for the IntentRadar system ($\mu = 3.5$) and for the IntentList system ($\mu = 3.5$) (two-sided Wilcoxon rank sum test with Bonferroni corrected Alpha, $W = 351$,

Question	IR	TQ	IL
This system provides adequate way to express preferences	3.6	3.2	3.7
This system provides adequate support to revise preferences	3.4	3.1	3.3
This system helps me to understand why the suggested articles should be important	2.7	2.8	3.5
The information provided by the system is sufficient to make decisions	3.5	3.0	3.1
The labels/keywords/information provided by the system are clear	4.0	3.2	4.0
The layout of the system is clear	2.8	2.7	2.5
I learned to use the system quickly	3.7	3.9	4.2
It did not take too much effort to find useful articles	2.5	2.3	3.0
I found it easy to express information need and preferences	3.5	3.0	3.2
I found it easy to train the system with updated preferences	2.3	2.3	2.3
With this system it is easy to alter the outcome of results	**3.6**	2.3	**3.4**
It is easy to get new set of items instead of what I already have	2.6	1.6	3.0
The system offered me useful options and avoided me getting stuck	3.2	2.8	3.3
I found it easy to explore the related areas without getting stuck	**2.9**	1.8	2.1
I feel in control to tell what I want	3.3	3.5	3.5
The system helps me to understand and keep track of why the items were relevant and offered for me	3.0	3.6	4.0
I'm satisfied with the system	3.6	3.4	3.1
I am convinced that I found the right articles	**2.6**	3.6	3.0
I would like to use the system, if offered for me	3.7	3.5	3.5
With this system it is easy to find answers to my information needs	2.8	2.4	3
I was able to take advantage of the system easily	3.5	3.5	3.7
The system influenced my choice of items	**4.1**	2.4	3.5
Mean of all questions	**3.2**	2.9	**3.2**

Table 4. Results of the selected ResQue questions. The mean user experience based on the ResQue questionnaire was found to be significantly higher for the systems (IR and IL) that use the intent modeling based interaction. The significantly higher values after correcting for multiple comparisons for each question are in bold face (IR denotes IntentRadar, IL denotes IntentList, and TQ denotes the Typed Query baseline).

$p < 0.05$) when compared to the Typed Query baseline system ($\mu = 3.2$). No differences between the IntentList and the IntentRadar Systems were found (two-sided Wilcoxon rank sum test with Bonferroni corrected Alpha, $W = 233.5$, $p = 1.0$).

The individual questions also revealed differences in participants' subjective preferences between the Typed Query baseline and the systems with interactive intent modeling, but differences did not emerge between the IntentList and the IntentRadar systems.

Both systems with interactive intent modeling assisted the participants to more adequately express their preferences and easier to alter the outcome of the results. The IntentRadar interface was found to have significantly better support for exploration of related areas without getting stuck and the participants felt that the system influenced their choice of items. Interestingly, the participants who used the IntentRadar system were significantly less convinced that they had found the right articles during the task. Given that the retrieval effectiveness was found to be significantly better for the IntentRadar system, and therefore the responses from the system were of better quality, a possible explanation is that because of the visualization the participants became more aware of other potentially relevant directions that they could not explore in the given time, and therefore might have been more informed about potentially relevant, but not yet explored, topics. Another possible explanation is that since the participants had little experience with the system, it may have increased their caution of additional, yet immediately not obvious information [33]; simply having more information might have raised doubts about whether better answers exist.

6 INFORMATION COMPREHENSION EXPERIMENT

The findings from the exploratory search experiment show that interactive intent modeling significantly improves retrieval performance over the search session. Simultaneously, however, participants' task performance is only improved when using the IntentRadar visualization and participants' subjective usability responses suggest that the IntentRadar visualization may have broader benefits in comprehending the retrieved information rather than just providing adequate feedback to direct the search.

Motivated by these findings, the IntentRadar visualization was studied in an additional experiment to reveal the effect of the visualization component to the user's information comprehension performance.

Whereas the exploratory search experiment evaluated performance over a search session, the information comprehension experiment aims to evaluate how well participants can comprehend a momentary set of search results presented to them.

In this experiment the focus is therefore not on feedback capabilities of the systems. Instead, in this experiment predefined queries will be presented to participants to allow comparing comprehension of the same search result sets across different systems. This allows decoupling the comprehension of a momentary search result set from how the results can be changed via feedback and to remove a confounding factor of which search result sets different participants would see.

6.1 Research Questions

In detail, the focus of the study was threefold. First, to study if the visualization would assist users in the comprehension process. Second, to study if the users preferred interaction with the visualization. Third, if the visualization would improve the output of the comprehension process. The experiment sought answers to the following research questions:

RQ5 Comprehension process: Do participants in the visualization condition inspect the search result space using the visualization more often than using the result list?

RQ6 Interaction support: Do participants in the visualization condition select keywords from the visualization more often than from the result list?

RQ7 Comprehension outcome: Does the visualization result in improved information comprehension outcome?

RQ8 User Experience: Does the result presentation using the visualization result in improved user experience?

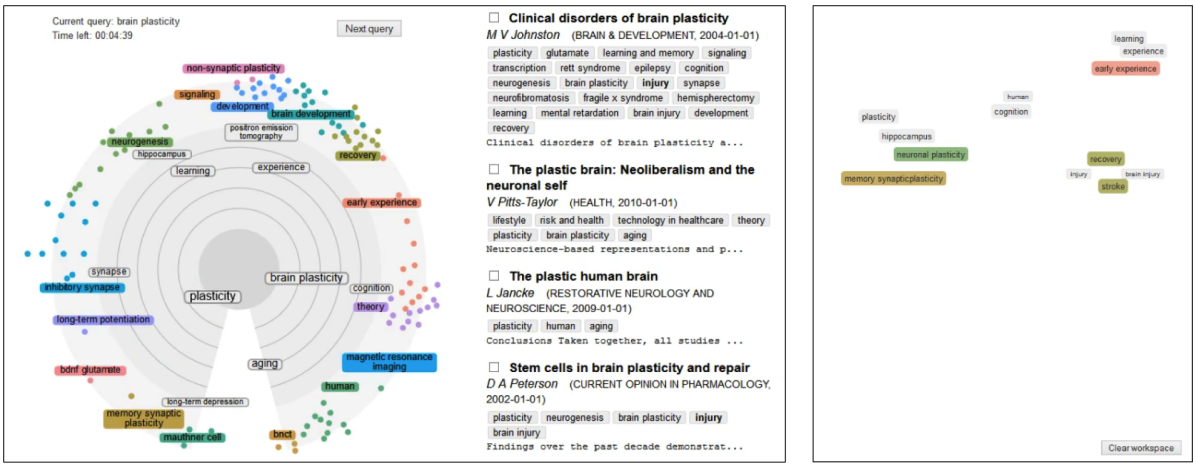

Fig. 8. A screenshot of the user interface that was used by the participants. The visualization component and the ranked list of search results (left) and the workspace (right). In the experiment, the workspace was placed under the visualization component as a floating element to ensure equal screen estate with the Typed Query baseline system.

6.2 Experimental Design

The independent variable of the experiment was the system configuration: a system with the visualization component and a system without the visualization component. We chose a 2×8 within-subjects design with two system configurations and eight topics. This design was chosen as tasks were short repetitive comprehension tasks and we wanted to avoid the cognitive effects due to the differences of participants. The ordering of the system conditions and the topics were counter-balanced and rotated using a Latin square design.

6.3 System Configurations

Two systems were configured: a system with the visualization component and a system without the visualization component. The system with the visualization component is illustrated in Figure 8. The system without the visualization component is exactly the same, but pertains only the conventional search result listing and the visualization component is removed (Figure 9). Both systems had the keywords visualized within search results (i.e. a list of keywords describing each document was placed under each document appearing in the result list). The systems were augmented with a workspace that was used by the participants to collect the information. This allowed a simple interaction to select information by dragging from the actual interface without switching to another application. The workspace also enabled accurate logging and data collection.

6.4 Task and Topics

The participants were situated in a simulated work task in which they had to comprehend and summarize the search results. The participants were asked to use two-level hierarchical conceptualization:

(1) Find as many *main topic keywords*, but at least two, that you find important to cover the overall topic.
(2) Find as many *subtopic keywords* under each main keyword that you find important to cover the main keyword.

The work task scenario was: "You are searching information about a pre-defined topic using an information retrieval system. Your task is to comprehend the topic by describing, at least two, main keywords related to the overall topic and describe as many as possible subconcepts by selecting keywords under the main keywords related to each main keyword." Eight topics were used: Human Memory, Web Design, Cognition, Distributed Systems, Language Processing, Kernel Function, Wearable Sensors, and Compiler Design.

6.5 Participants

We recruited 24 participants from two universities. Six were females. The participants were 20-40 years old. All the participants were graduate students with a technical background. As the text in the user interface was in English, only participants with a self-reported good knowledge of English were eligible to take part. Participants were told they could ask the experimenter for clarification at any time during the experiment. All participants had experience with interactive search engines, but participants were not familiar with any of the systems used in the experiments. Users were recruited by word of mouth and received no compensation for participation.

6.6 Procedure

Prior to the experiment, the participants were asked to read written instructions. The instructions explained the purpose of the experiment and the task that the participants were expected to perform. The participants were then informed that the system would automatically launch queries and return and present search results, and the participant was only expected to gather information by examining the presented search results using the given system. The participants were informed that they will use two different *systems* to gather information and store their conceptualization in a *workspace* component, which will be the same for both systems. Then the participants watched a 2-minute video illustrating an exemplar task with the system variants and they had a 3-minute trial using the systems via a pre-defined query that was different from the ones used in the actual experiment.

After this phase, the actual experiment started. Participants performed eight tasks corresponding to the eight topics. Each task had two phases: comprehension phase and composition phase.

The rationale of the comprehension phase was to study the quality of the keywords that the participants were able to produce given a system variant. The rationale in separating the composition phase was to let the users concentrate in producing the conceptualization as fast as possible and to allow them to organize the selected keywords in a separate composition phase. Previous research has shown that humans often spend significant amount of time in composing their answers instead of looking for information to support their answers and that these two tasks are interleaved and cause task-switching costs [80]. The separation of the phases was important as the participants had a strictly limited time to perform the comprehension phase and the output of that phase may have been interfered with the answer composition when performed under restricted time. The experimental design where these tasks were separated ensured that the participants focused on collecting the best possible keywords comprehending the result space in the given time without having to interleave this activity with composition of their answer.

In the comprehension phase, the system automatically launched queries corresponding to the tasks. Each query was run either on timer (5 minutes), or when the participant clicked the "next" button. The query that was automatically issued by the system was exactly the name of the topic. For example, for the topic *Web Design*, the system automatically issued a query "Web design". This allowed us to remove possible variance originating from participants' subjective interpretations of the topics. The time limit was used to make sure that there was no variance in the time that the participants used in the comprehension phase. The timer was visible for the

participants so that they were aware of how much time they had left to complete the task. The participants had two minutes to read and collect the information on the screen and after two minutes all the information on the screen disappeared, except the information on the workspace which was used to store the results to be used in the composition phase.

In the composition phase, the participants could still use the information that they had collected to the *workspace* to compose a written answer that comprehended the search result space. The workspace was visible for additional 3 minutes. Such experimental procedure ensured that the participants were working under strict time limits in order to complete the task as fast as possible and use the preferred interface element when they knew that their time is limited. After the experiment, the participants filled in post-task questionnaires selected from the ResQue questionnaires [86].

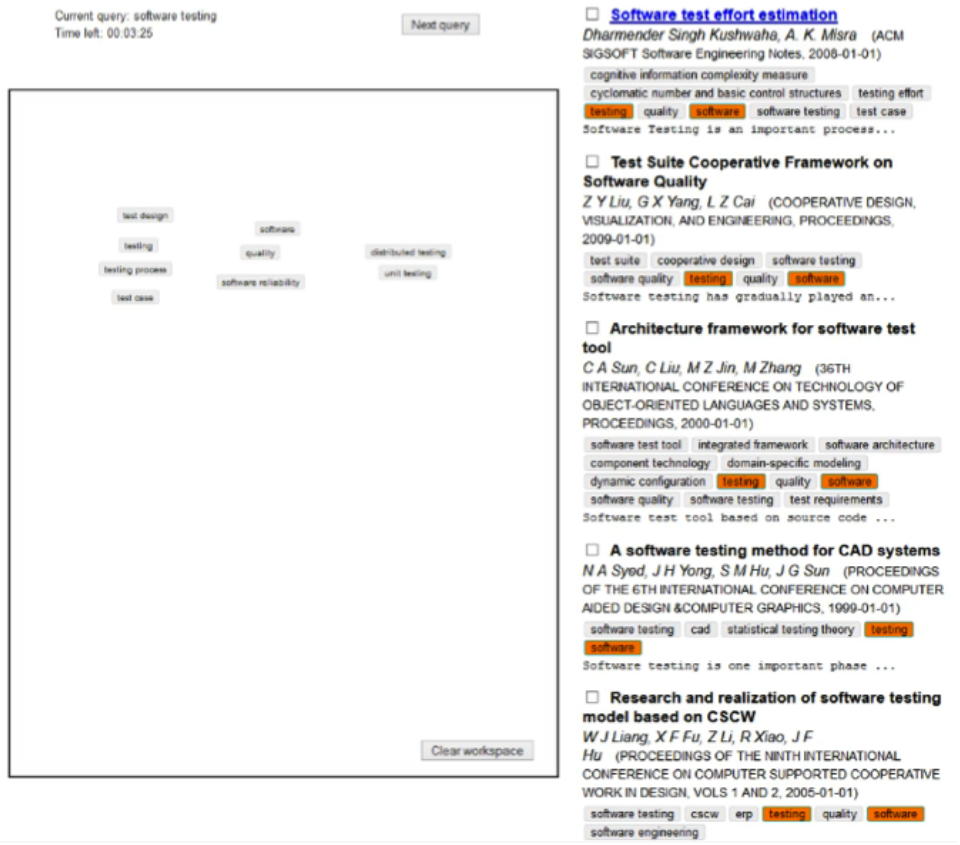

Fig. 9. A screenshot of the Typed Query baseline system without the visualization component. The ranked list of search results (right) and the workspace (left).

6.7 Apparatus

The experiment was run on a standard desktop PC connected to a vertically mounted 24-inch wide-screen monitor. The vertical position of the monitor was chosen because the workspace was placed under the result list and visualization component and the additional screen estate allowed fair comparison to the Typed Query

baseline system. The system was implemented as a Web application accessed using the Google Chrome browser. During the experiment, participants could use a mouse and a keyboard to operate the interface. The search engine automatically logged the timestamp and the action performed by the user. The recorded actions were: selection of a keyword to represent the main or subtopic, the position of the documents in the ranked list that contained the particular keyword, and the state of the workspace.

6.8 Data and Settings

We used the same data, indexing, and ranking models with the same settings as in the exploratory search experiment described in Section 4.8.

6.9 Relevance Assessments

After the experiment all responses from all participants and systems were pooled so that each *main topic keyword* and each *subtopic keyword* associated with the main topic keyword were listed in a matrix. Two assessors assessed the relevance of the main topic keywords and the subtopic keywords using a graded relevance on a 5-point Likert scale:

(1) Relevance of main topic — does this keyword represent an relevant overall area for the task topic? (very relevant (4), relevant (3), moderately relevant (2), somewhat relevant (1), irrelevant (0))
(2) Relevance of the subtopic —is this keyword relevant for the main topic? (very relevant (4), relevant (3), moderately relevant (2), somewhat relevant (1), irrelevant (0))

Essentially the assessment provided a goodness measure of each subtopic keyword and main topic keyword per task topic by removing overlapping instances. The scores were created by one expert and checked by another, resolving disagreements by consensus. Some topics were ambiguous because they had several interpretations. For example, the topic "language processing" had interpretation in computer science and human language processing. The experts also evaluated ambiguous topics accordingly; more comprehensive response would include the different interpretations of the concepts in the task.

6.10 Measures

The quality of comprehension process (RQ5) was measured as the inspection source detected from the mouse positions of the participants when inspecting the user interface. In particular we quantified the share of information inspection source: the time spent browsing the visualization compared to the time spent browsing the result list. It has been shown that mouse positions are associated with attention [26] and attention can be associated with human information processing [103, 108].

Interaction support (RQ6) was measured as the frequency of the usage of the interaction elements. In particular we quantified the share of information selection source: the frequency of selection of keywords from the visualization compared to selection from the result list.

Comprehension outcome (RQ7) was measured as the cumulative gain of selected keywords.

The cumulative gain was computed as the sum of the relevance assessed for the keywords selected by the participants. The gain is computed separately over main topics and within each subtopic identified in the set of keywords. Thus, the cumulative gain quantifies both the main topic coverage associated with the diversity of the keywords and coverage within each subtopic associated with comprehensiveness of the keywords in each subtopic.

User experience (RQ8) was measured using the ResQue questionnaires: how participants subjectively rated the usefulness and usability of the compared systems.

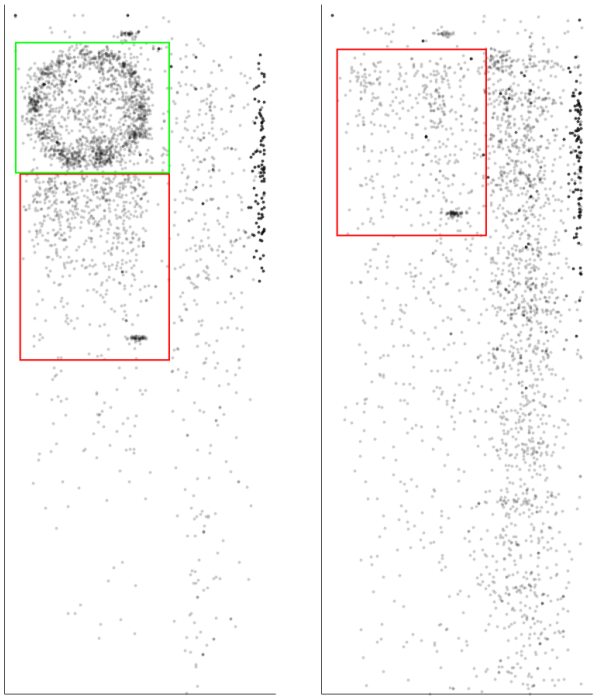

Fig. 10. Mouse position scatter plots from 18 users over the two systems: the one with the visualization (left) and the Typed Query baseline (right). Dots are mouse positions recorded at 3-second intervals. The areas of the workspace and the visualization are outlined in red and green respectively. Dots in the bottom-half of the figures are from situations where the user has scrolled the screen to see more results.

We next present the results of the above-described measurements, and the conclusions to the research questions RQ5-RQ8 will then be summarized in Section 8.2 along with the research questions RQ1-RQ4 of the exploratory search experiment.

7 RESULTS OF THE INFORMATION COMPREHENSION EXPERIMENT

7.1 Share of Information Inspection Source

Recorded mouse movements over time were available from 18 of the users (mouse-movements of 6 users were not available due to a technical problem). Figure 10 shows the locations as a scatter plot. Based on the mouse locations, users on the system with the visualization spent 30.5% of time browsing the map, and 55.4% of time browsing the search result list. In comparison, users on the Typed Query baseline system spent 83.4% of time browsing the search result list. This further illustrates the fluency of the visualization: users spent a reasonable portion of time browsing the map. While less time was spent over the map than over the result list, the majority of keywords were dragged from the map.

7.2 Share of Information Selection Source

Users of the Typed Query baseline system dragged in total 1286 keywords from under articles in the result list to the workspace, on average 6.7 per user and task. Users of the system with the visualization dragged in total 1068

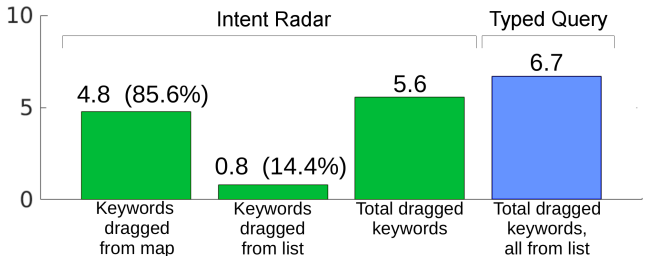

Fig. 11. Sources of information selected to be dragged to the workspace. Numbers are average amounts of keywords dragged to the workspace, for both systems, on average over all tasks and users. For the system with the visualization, we also report separately the amounts of keywords dragged from the map and from the document list, and give their relative percentages. On the system with the map, users drag from the map most of the time.

keywords to the workspace, on average 5.6 per user and task. The total amount of keywords selected was not found to be statistically significant (Welch Two Sample t-test, df = 44.842, p = 0.15). Participants in the IntentRadar condition strongly preferred to use the IntentRadar visualization and dragged on average 4.8 keywords from the visualization and only 0.8 keywords from the list. The difference was found to be statistically significant (Welch Two Sample t-test„df = 29.85, $p<0.00001$). All in all, on the system with the visualization, 89.8% of the keywords were dragged from the visualization. As we discuss in the next section, the cumulative gain of the selected keywords is higher for the IntentRadar system than for the Typed Query baseline; thus, even though users overall dragged slightly fewer keywords onto the workspace in the IntentRadar system than in the Typed Query baseline, the visualization allowed them to reach a more comprehensive selection of keywords than the Typed Query baseline system.

The effect of the IntentRadar substituting the list was confirmed by comparing the amount of keywords dragged from the list in the Typed Query baseline system and in the IntentRadar system (Welch Two Sample t-test, df = 32.766, $p<0.00001$). The result shows that the differences were because the keywords were not only selected from the IntentRadar as complementary to the list, but used as a substitute to the list. Figure 11 shows these results graphically.

7.3 Cumulative gain of selected keywords

Users had been asked to provide answers organized as main-topic keywords and sub-topic keywords under each main topic, and both were evaluated separately by cumulative gain of expert-given scores for the keywords. Average within-task standard deviation over users was 3.9 for main-topic scores and 17.2 for sub-topics; we focus on between-system difference. Figure 12 shows that the IntentRadar users yielded a statistically significantly improved score for main-topic keywords; since the main topics represent the breadth of information content discovered from the results, the visualization helped users reach a more comprehensive understanding of the results. Figure 13 shows the results for sub-topic keywords, representing depth of understanding for each main topic; the difference between systems was not statistically significant, hence the visualization increased overall comprehension without sacrificing depth of comprehension.

7.4 Subjective preference

In the post-task questionnaires (Table 5), users indicated through ratings of several questions that the simpler and more familiar interface was found easier to use and learn, and they felt more confident using the conventional system with only search result listing. This is natural when comparing a traditional interface to a new one with only a small amount of training time. However, users clearly felt the visualization influenced their selection of

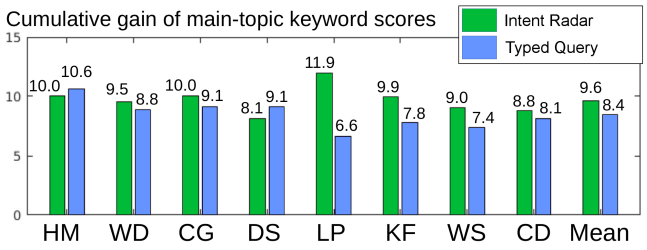

Fig. 12. Cumulative gain of main-topic keyword scores from experts. Numbers are cumulative-gain scores averaged over the users, for each task and each system. The tasks are: Human memory (HM), Web design (WD), Cognition (CG), Distributed systems (DS), Language processing (LP), Kernel functions (KF), Wearable sensors (WS), Compiler design (CD). The rightmost "Mean" bars are the mean over all tasks per system. The IntentRadar system is statistically significantly better than the IntentList system, by right-tailed two-sample t-test at the $p = 0.05$ threshold ($p = 0.0192$).

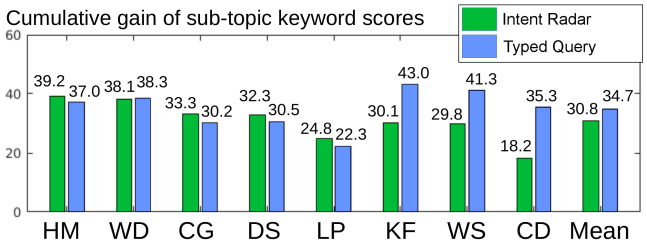

Fig. 13. Cumulative gain of sub-topic keyword scores from experts, with respect to their corresponding main-keywords. Numbers are cumulative-gain scores averaged over the users, for each task and each system. The tasks are the same as in Figure 12, and the rightmost bars are the mean over all tasks per system. The overall difference between the systems is not statistically significant.

topics and subtopics. While users' overall satisfaction score for the systems was similar, in a separate question about system preference, a two thirds majority of the 24 users preferred to use the system with the visualization.

8 DISCUSSION AND CONCLUSIONS

The current generation of information retrieval systems, such as the major Web search engines, is effective at identifying a small set of the most relevant documents given a well specified information need. However, it is easy to identify many situations where more complex exploratory search support is required and increasing real-world evidence suggests that users are struggling with exploratory search [81]. As a consequence, an important goal of an information retrieval system is to assist the user in understanding and specifying his/her information needs [50]. In particular, when users are engaged in complex tasks, search may not be best supported with simplistic search user interfaces and ranking optimized to maximize relevance to a single query, but rather users are engaged with the system to maximize whole-session relevance [90]. In exploratory search, what is encountered along the exploration affects the search intents and goals. Consequently, the system must provide the user with affordances to comprehend the information space, direct the search, and engage the user in the exploration process. At the same time, the system can learn from user interactions to assist the user in accomplishing the user's task. Achieving a common understanding about search intents and goals requires an intrinsic interplay between the user and the information retrieval system.

Question	IR	TQ	p-value
I found the system unnecessarily complex	2.5	**1.7**	0.004
I thought the system was easy to use	3.8	**4.4**	0.002
I think that I would need the support of a technical person to be able to use this system	1.9	**1.5**	0.03
I found the various functions in this system were well integrated	3.1	**3.6**	0.03
I thought there was too much inconsistency in this system	2.8	**1.9**	0.0002
I would imagine that most people would learn to use this system very quickly	3.3	**4.2**	0.003
I found the system very cumbersome to use	2.9	**2.1**	0.04
I felt very confident using the system	3.0	**3.8**	0.01
I needed to learn a lot of things before I could get going with the system	2.3	**1.7**	0.03
The system can be trusted	3.0	**3.8**	0.003
I became familiar with the system very quickly	3.7	**4.4**	0.008
The labels/keywords/information provided by the system are clear	3.0	**3.7**	0.03
The system influenced my selection of topics and subtopics	**4.1**	3.0	$2 \cdot 10^{-4}$
Which system do you prefer?	**16**	8	

Table 5. Post-task questionnaires, selected from the ResQue questionnaires, in which significant differences were found. Numbers are 5-point Likert scale agreement scores (Strongly disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly agree (5)). with the statements in the question column, averaged over the 24 users for each system: IntentRadar (IR), Typed Query baseline (TQ), and the t-test p-value of the difference. The better score for each question is in bold. The last line is the question of system preference, where we directly list how many users preferred each system; 67% preferred the IntentRadar.

8.1 Methodological Contributions

We contributed the principle of interactive intent modeling for exploratory search, and demonstrated the technique as a part of a real interactive information retrieval system. Interactive intent modeling allows interactive modeling of user's information needs and diversified presentation of concepts that the user can utilize as exploration affordances. This allows improved communication between the information retrieval system and the user. As a result, the user can learn about the available exploration possibilities and comprehend the potential directions around the user's present position in the information space. Conversely, the system can adapt to the user's evolving intentions that arise during the exploration process.

In our approach, the present and potential future search intents are estimated at the task level. The estimation is based both on task context and on user interaction that rewards or penalizes the intent model implemented as a multi-armed bandit reinforcement learning system. In this way, the intent modeling allows the user to retrieve

information relevant for the present search intent estimate and to interact with anticipated search intents that help in reducing the system's uncertainty related to the user's evolving intents. The intent model is visualized for information comprehension and interaction to allow the user to direct the search. Unlike the conventional relevance feedback approach, the key idea is to estimate the relevances of keywords using the model that can consist of information beyond what is present at the top-ranked results. Our technique also demonstrates how the visualization of the model can turn the human memory recall task into a fluid visual recognition task, that in turn can assist the user in expressing intents by interacting with the visual interfaces.

Some recent information seeking research on task-level relevance and task performance point to similar results as we reported. For example, improved task performance has been associated with improvements in recall despite higher search efforts or degrading precision [119]. In complex tasks, users have also been shown to have trouble in finding and assessing relevance of novel information [107] and that successful search sessions tend to involve more user effort than unsuccessful sessions [93]. This is in line with our findings showing that more engagement and search effort seems to lead to improved retrieval performance over the session and improved task outcomes. Recent research has also found exploration/exploitation to be effective in acquiring novel information [52]. This supports our findings that rigorously modeling relevance and uncertainty, while keeping humans in control, even with a tradeoff of increased interaction and momentary decrease in precision, are key factors in succeeding in exploratory search.

We reported on two experiments, studying exploratory search support and information comprehension support. Next we summarize and reflect our findings in light of the research questions.

8.2 Empirical Evidence

The exploratory search experiment and the information comprehension experiment were carried out to answer the specific research questions that are reflected below.

RQ1 Task performance in the exploratory search task: Does interactive intent modeling lead to better task outcome? Yes, but the improvements are dependent on combining the intent modeling with a visualization that helps comprehending the search results (Section 5.1).

RQ2 Retrieval performance in the exploratory search task: Does interactive intent modeling result in high-quality retrieved information? Yes, the improvements can be attributed to recall of novel information over the session, while sustaining the precision and recall of novel and obvious information (Sections 5.2 and 5.3). However, the retrieval results only transfer to improved task performance when they are visualized appropriately.

RQ3 Interaction support in the exploratory search task: Does interactive intent modeling elicit useful interactions? Yes, the systems with interactive intent modeling elicited significantly more interaction (Section 5.4) that in turn transferred into improved retrieval performance. However, the interactions with the intent model did not replace, but rather complemented typed query interaction.

RQ4 User experience in the exploratory search task: Does the increased complexity of the user interface design, compared to standard search interfaces, affect the subjective user experience? Yes, participants found support for influencing their search behavior, altering the outcome of the system, and exploring without getting stuck (Section 5.5). However, the system also raised participants' concerns whether they found the right results.

RQ5 Comprehension process in the search result comprehension task: Do the participants in the visualization condition inspect the search result space using the visualization more often than using the result list? Yes, the participants, in the visualization condition, inspected the search results using the visualization. The recordings of the mouse movements (Section 7.1) showed that over one third of the time the users inspected the search results using the visualization. However, the visualization did not replace the conventional result list as one third of the time was spent inspecting the result list.

RQ6 Interaction support in the search result comprehension task: Do the participants in the visualization condition select keywords from the visualization more often than from the result list? Yes, participants, in the visualization condition, selected a large majority (89.8%) of their dragged keywords from the visualization rather than from the result list (Section 7.2).

RQ7 Comprehension outcome in the search result comprehension task: Does the visualization improve information comprehension outcome? Yes, the visualization improved the comprehension outcome, as measured by topic feature coverage, for the main topic keywords (Section 7.3). However, differences were not found in the case of the subtopic keywords. This suggests that the visualization enabled participants to obtain a broader view on the search results, but did not help gather better information under an individual subtopic.

RQ8 User experience in the search result comprehension task: Does the result presentation using the visualization result in improved user experience? No, in the search result comprehension task the participants felt that the visualization was in general not improving their user experience (Section 7.4), and were more confident when using the conventional system wih only a search result listing. However, when asked which user interface they would prefer, a significantly larger portion of the participants preferred the visualization over the conventional system with only the result list. This might suggest the participants found the better comprehension enabled by the visualization condition was worth any additional complexity and lesser confidence in the user experience.

8.3 Implications

The implications of the results for user modeling and exploratory search interfaces are significant because they open opportunities to learn user models from natural user interactions by visualizing the the task-level knowledge model representing the present and potential search intentions of a user and allowing relevance feedback directly on the model. This has direct implications for designing exploratory search systems that can be summarized as follows.

Implication 1: interactive intent modeling improves retrieval performance. Our results indicate that the retrieval performance of the system is improved as a result of interactive intent modeling. Retrieval performance is not dependent on the type of visualization and can be attributed to the intent modeling.

Implication 2: Retrieval performance transfers to task performance when the intent model is visualized for effective comprehension of the information space. Users' task performance is dependent on how they can comprehend the retrieved information and improved retrieval performance is only transferred to improvements in task performance with the IntentRadar visualization. This implies that while reduced visualizations or other query augmentation techniques may be sufficient for improving document retrieval performance, the benefits may not transfer to task performance in exploratory search tasks.

Our results suggest that exploratory search can benefit from visual search support. Support for concept recovery and mental work in the form of visually-supported reasoning in the visual space suggest opportunities for improved task outcomes and improved retrieval performance over the session. The cognitive demands of exploratory search are possibly associated with a preference of recognition over memory recall for users present domain knowledge and reflecting new information confronted during exploration for assisting users to move towards their task goals [46]. These benefits can be operationalized by externalizing memory and knowledge representation that can be particularly beneficial in exploratory search scenarios in which users are investing substantial time and cognitive resources for reaching their goals [115].

Implication 3: Interaction with the intent model complements typed query interaction. The interactions with the intent model were found to be complementary to the conventional typed query interaction. Participants used the interactive intent model to direct the search, while typed query interaction was used to "teleport" to a new area in the information space. Participants also subjectively reported that the system influenced

their search behavior and allowed them to express and revise search preferences. Visualizing more information may be associated with increased user effort and the users may exhibit increased scanning times and difficulty of expressing their intentions. Our experiments, however, suggest that interactive visualization of search intentions improves user performance without compromising search effort.

Implication 4: Optimized intent model visualization improves search result comprehension, but does not increase user effort. The interactions with the intent model did not increase user effort in terms of interaction time, but participants interacted even faster with complex visualizations and visualizations were found to assist the participants in obtaining an improved information coverage of the search result space. The participants also voluntarily spent most of the time investigating the visualization instead of the conventional search result list. The subjective feedback indicated that the systems positively affected the participants' search behavior. Participants also reported that they were less certain that they found all the right results. These findings suggest that visualized models can provide the users affordances, not only to direct their search, but also to make sense of the information potentially available. The users also take advantage of these affordances without compromising the time spent between interactions.

Implication 5: Direct visual manipulation of the modeled intent is beneficial for exploratory search. Our results imply that supporting exploratory search behavior requires support for both information comprehension and search directing. The experiments suggest that these are best conveyed to the users by using the visualization for two means: allowing users to perceive and comprehend the information space, but also to directly manipulate the model to provide feedback on their evolving intentions while they reflect the retrieved information. Our results also show that increased information on the IntentRadar visualization allows effective and efficient direct manipulation of the intent model without compromising time to make decisions. We believe it is evidence that users are able to process the information presented to them and provide effective feedback to the system. Together with the improved results in task and retrieval performance, fast and effective interaction shows that user can make efficient and effective decisions, and that these decisions are beneficial for their task performance.

8.4 Limitations and Future Work

The specific limitation of the chosen experiments was the sole focus on aspectual exploratory search tasks. Participants were given simulated work tasks to cover multiple aspects of a topic. The effects and results are therefore limited to aspectual exploratory search settings and the benefits of interactive intent modeling should be interpreted in this context.

Another limitation is the selection of the factors that were studied in the user experiments. For example, while we implemented different system variants and interactive visualizations, some other visualization techniques may be equally effective and engaging. Moreover, users might make increased use of the visualization elements by quick glances that are not revealed in mouse movements. While the mouse logs provide evidence for attention allocation, eye tracking instrumentation would provide more accurate information and could be used to confirm the proportion of attention allocation in the search support spaces. Similarly, the present experimental setup did not separate different cognitive states that the users may experience, such as scanning, deciding, hesitation, or confirming. Associating more precise cognitive states to user behavior with intent modeling and search result visualization are interesting future endeavours to increase our understanding on the specific type of user support.

The experiments were designed to minimize confounding factors and biases caused by task and system ordering, as well as domain-knowledge effects. However, as the IntentRadar system contains novel visualization, a novelty bias towards the system with the visualization is possible; the participants may have focused their attention to the novel visualization partly because it is novel – an effect that may have been less strong when participants would have used the system for a longer period. The participants in the experiments were screened for low and

high pre-knowledge. However, the possibility of domain-knowledge effects are hard to exclude completely and their association with the utility of the visualization, and intent modeling in general, remains an interesting future research direction.

Different approaches to balance exploitation and exploration could also be studied in principle, and an extensive simulation could be conducted to study different tradeoffs of such models. The present experiments were limited to aspectual search tasks and future work should be conducted to generalize an intent model algorithm over the range of complex exploratory tasks. These could investigate combining the reinforcement learning approach with diversification, modeling task stages, exploratory tasks that have distinctive contextual structures, for example searching and combining information simultaneously from distinct areas or their intersections, and automatically learning exploration/exploitation tradeoffs suitable for different types of tasks and phases within tasks.

In the present experiments, we did not intend to study all possible interaction means that could be deployed in search systems. Future work could compare implementations of interactive intent modeling within other types of search user interfaces, such as parts of query suggestion or query autocompletion interfaces, or the modeling technique in combination of faceted search or interactive tag clouds. These are potentially important aspects when considering deployment in real-life settings beyond tasks studied in our experiments.

Moreover, experiments that would allow capturing more precise user signals for studying users' attention allocation, such as eye tracking instrumentation, could allow confirming the proportion of attention allocation in the search support spaces to refine conclusions. It is possible that participants make increased use of the visualization elements by quick glances that are not revealed in mouse movements as measured in our experiments. Mouse movements also require motor control planning and may therefore be only partially indicative of the user's attention targets when a user is considering a visual element but not fully committed to an interaction.

However, the principle of interactive intent modeling can enable the user to interact with the intent model through a visualization in order to provide feedback directly on the model. The technical realization of the approach by using reinforcement learning with rewards obtained from user interactions and the effectiveness of the intent model for retrieval and the visualization for search result comprehension can be applied to information seeking beyond the present studies, for example information exploration on the Web.

ACKNOWLEDGMENTS

This work has been partly supported by the Academy of Finland (278090; 252845; 305739; 252845;294238; 292334; 255725; and the Finnish Centre of Excellence in Computational Inference Research COIN), Re:Know and D2I funded by TEKES, and MindSee (FP7 ICT; Grant Agreement 611570). Certain data included herein are derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011. All rights reserved. Data is also included from the Digital Library of the ACM, the Digital Library of IEEE, and the Digital Library of Springer. We want to thank all researchers participating in HWFA, Multivire, D2I and Re:Know for their help and assistance.

REFERENCES

[1] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennet. 2012. Search, Interrupted: Understanding and Predicting Search Task Continuation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 315–324. https://doi.org/10.1145/2348283.2348328

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/1498759.1498766

[3] Christopher Ahlberg and Ben Shneiderman. 1994. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 313–317. https://doi.org/10.1145/191666.191775

[4] Jae-wook Ahn and Peter Brusilovsky. 2009. Adaptive visualization of search results: Bringing user models to visual analytics. *Information Visualization* 8, 3 (2009), 167–179.

[5] Jae-wook Ahn and Peter Brusilovsky. 2013. Adaptive Visualization for Exploratory Information Retrieval. *Information Processing & Management* 49, 5 (2013), 1139–1164. https://doi.org/10.1016/j.ipm.2013.01.007

[6] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. 2007. Open User Profiles for Adaptive News Systems: Help or Harm?. In *Proceedings of the 16th international conference on World Wide Web*. ACM, ACM, New York, NY, USA, 11–20.

[7] Jae-wook Ahn, Peter Brusilovsky, Daqing He, Jonathan Grady, and Qi Li. 2008. Personalized Web Exploration with Task Models. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, ACM, New York, NY, USA, 1–10.

[8] Jae-wook Ahn, Peter Brusilovsky, and Shuguang Han. 2015. Personalized Search: Reconsidering the Value of Open User Models. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 202–212. https://doi.org/10.1145/2678025.2701410

[9] Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. 2007. Exploratory Search Using Timelines. In *Proceedings of the ACM SIGCHI 2007 Workshop on Exploratory Search and HCI*. ACM, New York, NY, USA, 23–26.

[10] John R. Anderson. 2000. *Learning and Memory: An Integrated Approach* (1st ed.). John Wiley & Sons, New Jersey, USA.

[11] Peter Auer. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research* 3 (2002), 397–422.

[12] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2004. Query Recommendation Using Query Logs in Search Engines. In *Proceedings of the 2004 International Conference on Current Trends in Database Technology*. Springer, Berlin, Heidelberg, 588–596. https://doi.org/10.1007/978-3-540-30192-9_58

[13] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/2449396.2449405

[14] Marcia J. Bates. 1986. Subject access in online catalogs: A design model. *Journal of the American Society for Information Science* 37, 6 (1986), 357–376.

[15] Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–431.

[16] Marcia J. Bates. 2007. What is Browsing–Really? A Model Drawing from Behavioural Science Research. *Information Research* 12, 4 (2007), paper 330.

[17] N. J. Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5 (1980), 133–143.

[18] N. J. Belkin, R. N. Oddy, and H. M. Brooks. 1982. Ask for Information Retrieval: Part I.: Background and Theory. *Journal of Documentation* 38, 2 (1982), 61–71.

[19] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, ACM, New York, NY, USA, 185–194. https://doi.org/10.1145/2348283.2348312

[20] John Brooke. 1996. SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189 (1996), 194.

[21] Katriina Byström and Kalervo Järvelin. 1995. Task Complexity Affects Information Seeking and Use. *Information Processing & Management* 31, 2 (1995), 191–213. https://doi.org/10.1016/0306-4573(95)80035-R

[22] Fei Cai, Ridho Reinanda, and Maarten De Rijke. 2016. Diversifying Query Auto-Completion. *ACM Transactions on Information Systems* 34, 4, Article 25 (June 2016), 33 pages. https://doi.org/10.1145/2910579

[23] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. 2009. Context-aware Query Classification. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/1571941.1571945

[24] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. Overview of the TREC 2014 Session Track. In *TREC'14*.

[25] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. 2011. Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 167–176. https://doi.org/10.1145/1978942.1978967

[26] Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. 2001. What Can a Mouse Cursor Tell Us More?: Correlation of Eye/Mouse Movements on Web Browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01)*. ACM, New York, NY, USA, 281–282. https://doi.org/10.1145/634067.634234

[27] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2017. Personalized Query Suggestion Diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 817–820. https://doi.org/10.1145/3077136.3080652

[28] Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively Predicting Diverse Search Intent from User Browsing Behaviors. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1772690.1772714

[29] Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively Predicting Diverse Search Intent from User Browsing Behaviors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1772690.1772714

[30] Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence Clustering and Labeling for Unsupervised Query Intent Discovery. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 383–392. https://doi.org/10.1145/2124295.2124342

[31] Paul-Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. 2007. Personalized Query Expansion for the Web. In *Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 7–14.

[32] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User Activity Patterns During Information Search. *ACM Transactions on Information Systems* 33, 1, Article 1 (March 2015), 39 pages. https://doi.org/10.1145/2699656

[33] Michael J. Cole, Xiangmin Zhang, Chang Liu, Nicholas J. Belkin, and Jacek Gwizdka. 2011. Knowledge Effects on Document Selection in Search Results Pages. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 1219–1220. https://doi.org/10.1145/2009916.2010128

[34] Roy Davies. 1989. The Creation of New Knowledge by Information Retrieval and Classification. *The Journal of Documentation* 45, 4 (1989), 273–301.

[35] Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank As You Go: User-Driven Exploration of Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 118–129. https://doi.org/10.1145/2856767.2856797

[36] Marian Dörk, Carey Williamson, and Sheelagh Carpendale. 2012. Navigating Tomorrow's Web: From Searching and Browsing to Visual Exploration. *ACM Transactions on the Web* 6, 3, Article 13 (Oct. 2012), 28 pages. https://doi.org/10.1145/2344416.2344420

[37] Geoffrey M. Draper, Yarden Livnat, and Richard F. Riesenfeld. 2009. A Survey of Radial Methods for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 5 (2009), 759–776. https://doi.org/10.1109/TVCG.2009.23

[38] Marina Drosou and Evaggelia Pitoura. 2010. Search result diversification. *ACM SIGMOD Record* 39, 1 (2010), 41–47.

[39] Huizhong Duan and ChengXiang Zhai. 2015. Mining Coordinated Intent Representation for Entity Search and Recommendation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 333–342. https://doi.org/10.1145/2806416.2806557

[40] Carsten Eickhoff, Kevyn Collins-Thompson, Paul N. Bennett, and Susan Dumais. 2013. Personalizing Atypical Web Search Sessions. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*. ACM, ACM, New York, NY, USA, 285–294. https://doi.org/10.1145/2433396.2433434

[41] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2556195.2556217

[42] Rafael Glater, Rodrygo L.T. Santos, and Nivio Ziviani. 2017. Intent-Aware Semantic Query Annotation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 485–494. https://doi.org/10.1145/3077136.3080825

[43] Dorota Glowacka, Tuukka Ruotsalo, Ksenia Konyushkova, Kumaripaba Athukorala, Giulio Jacucci, and Samuel Kaski. 2013. Directing Exploratory Search: Reinforcement Learning from User Interactions with Keywords. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 117–128. https://doi.org/10.1145/2449396.2449413

[44] Dorota Glowacka and John Shawe-Taylor. 2010. Content-based image retrieval with multinomial relevance feedback. In *Proceedings of 2nd Asian Conference on Machine Learning*. 111–125.

[45] Qi Guo and Eugene Agichtein. 2009. Beyond Session Segmentation: Predicting Changes in Search Intent with Client-Side User Interactions. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 636–637. https://doi.org/10.1145/1571941.1572053

[46] Jacek Gwizdka and Michael Cole. 2013. Does Interactive Search Results Overview Help?: An Eye Tracking Study. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 1869–1874. https://doi.org/10.1145/2468356.2468691

[47] Susan Havre, Elizabeth Hetzler, Ken Perrine, Elizabeth Jurrus, and Nancy Miller. 2001. Interactive Visualization of Multiple Query Results. In *Proceedings of the IEEE Symposium on Information Visualization 2001*. IEEE Computer Society, Washington, DC, USA, 105.

[48] Daqing He, Peter Brusilovsky, Jae-wook Ahn, Jonathan Grady, Rosta Farzan, Yefei Peng, Yiming Yang, and Monica Rogati. 2008. An Evaluation of Adaptive Filtering in the Context of Realistic Task-Based Information Exploration. *Information Processing & Management* 44, 2 (2008), 511–533.

[49] Marti A. Hearst. 1995. TileBars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 59–66. https://doi.org/10.1145/223904.223912

[50] Marti A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press, New York, NY, USA. http://searchuserinterfaces.com/book/

[51] Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 76–84. https://doi.org/10.1145/243199.243216

[52] Katja Hofmann, Shimon Whiteson, and Maarten Rijke. 2013. Balancing Exploration and Exploitation in Listwise and Pairwise Online Learning to Rank for Information Retrieval. *Information Retrieval* 16, 1 (Feb. 2013), 63–90. https://doi.org/10.1007/s10791-012-9197-9

[53] Botao Hu, Yuchen Zhang, Weizhu Chen, Gang Wang, and Qiang Yang. 2011. Characterizing Search Intent Diversity into Click Models. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 17–26. https://doi.org/10.1145/1963405.1963412

[54] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search Result Diversification Based on Hierarchical Intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 63–72. https://doi.org/10.1145/2806416.2806455

[55] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. *Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions*. Springer Berlin Heidelberg, Berlin, Heidelberg, 4–15. https://doi.org/10.1007/978-3-540-78646-7_4

[56] Xiaoran Jin, Marc Sloan, and Jun Wang. 2013. Interactive Exploratory Search for Multi Page Search Results. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 655–666. https://doi.org/10.1145/2488388.2488446

[57] Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1972), 11–21.

[58] Mika Käki. 2005. Findex: Search Result Categories Help Users When Document Ranking Fails. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 131–140. https://doi.org/10.1145/1054972.1054991

[59] Evangelos Kanoulas, Ben Carterette, Paul Clough, and Mark Sanderson. 2010. Overview of the TREC 2010 Session Track. In *TREC'10*.

[60] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Investigating Users' Query Formulations for Cognitive Search Intents. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 577–586. https://doi.org/10.1145/2600428.2609566

[61] Dian Kelly and Xin Fu. 2006. Elicitation of Term Relevance Feedback: An Investigation of Term Source and Context. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 453–460. https://doi.org/10.1145/1148170.1148249

[62] Diane Kelly, Karl Gyllstrom, and Earl W. Bailey. 2009. A Comparison of Query and Term Suggestion Features for Interactive Searching. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 371–378. https://doi.org/10.1145/1571941.1572006

[63] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. Intent Models for Contextualising and Diversifying Query Suggestions. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 2303–2308. https://doi.org/10.1145/2505515.2505661

[64] Gary Klein, Brian Moon, and Robert R. Hoffman. 2006. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 21, 4 (July 2006), 70–73. https://doi.org/10.1109/MIS.2006.75

[65] Khalil Klouche, Tuukka Ruotsalo, Diogo Cabral, Salvatore Andolina, Andrea Bellucci, and Giulio Jacucci. 2015. Designing for Exploratory Search on Touch Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4189–4198. https://doi.org/10.1145/2702123.2702489

[66] Khalil Klouche, Tuukka Ruotsalo, Luana Micallef, Salvatore Andolina, and Giulio Jacucci. 2017. Visual Re-Ranking for Multi-Aspect Information Retrieval. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 57–66. https://doi.org/10.1145/3020165.3020174

[67] Weize Kong, Rui Li, Jie Luo, Aston Zhang, Yi Chang, and James Allan. 2015. Predicting Search Intent Based on Pre-Search Context. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 503–512. https://doi.org/10.1145/2766462.2767757

[68] Carol C. Kuhlthau. 2004. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited Westport, CT, Westport, CT, USA.

[69] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What Do Exploratory Searchers Look at in a Faceted Search Interface?. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*. ACM, New York, NY, USA, 313–322. https://doi.org/10.1145/1555400.1555452

[70] William Kules, Max L. Wilson, Monica C. Schraefel, and Ben Shneiderman. 2008. *From Keyword Search to Exploration: How Result Visualization Aids Discovery on the Web*. Technical Report. University of Southampton. http://eprints.soton.ac.uk/265169/

[71] Zhen Liao, Yang Song, Li-wei He, and Yalou Huang. 2012. Evaluating the Effectiveness of Search Task Trails. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 489–498. https://doi.org/10.1145/2187836.2187903

[72] Michael L. Littman. 2015. Reinforcement learning improves behaviour from evaluative feedback. *Nature* 521 (2015), 445–451.

[73] Jingjing Liu and Nicholas J. Belkin. 2010. Personalizing Information Retrieval for Multi-session Tasks: The Roles of Task Stage and Task Type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, New York, NY, USA, 26–33. https://doi.org/10.1145/1835449.1835457

[74] Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win Search: Dual-agent Stochastic Game in Session Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 587–596. https://doi.org/10.1145/2600428.2609629

[75] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[76] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2012. Citeology: Visualizing Paper Genealogy. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/2212776.2212796

[77] Alessandro Micarelli, Fabio Gasparetti, Filippo Sciarrone, and Susan Gauch. 2007. Personalized Search on the World Wide Web. In *The Adaptive Web*. Springer, Berlin, Heidelberg, 195–230.

[78] Matthew Mitsui, Jiqun Liu, Nicholas J. Belkin, and Chirag Shah. 2017. Predicting Information Seeking Intentions from Search Behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 1121–1124. https://doi.org/10.1145/3077136.3080737

[79] Matthew Mitsui, Chirag Shah, and Nicholas J. Belkin. 2016. Extracting Information Seeking Intentions for Web Search Sessions. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 841–844. https://doi.org/10.1145/2911451.2914746

[80] Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3 (2003), 134–140.

[81] Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. 2015. Struggling and Success in Web Search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1551–1560. https://doi.org/10.1145/2806416.2806488

[82] Jaakko Peltonen, Kseniia Belorustceva, and Tuukka Ruotsalo. 2017. Topic-Relevance Map: Visualization for Improving Search Result Comprehension. In *Proceedings of the 22nd International Conference on Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 611–622. https://doi.org/10.1145/3025171.3025223

[83] Peter Pirolli and Stuart Card. 1995. Information Foraging in Information Access Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 51–58. https://doi.org/10.1145/223904.223911

[84] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643.

[85] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM, New York, NY, USA, 275–281. https://doi.org/10.1145/290941.291008

[86] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the fifth ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[87] Filip Radlinski and Susan Dumais. 2006. Improving Personalized Web Search Using Result Diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 691–692. https://doi.org/10.1145/1148170.1148320

[88] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning Diverse Rankings with Multi-armed Bandits. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*. ACM, New York, NY, USA, 784–791. https://doi.org/10.1145/1390156.1390255

[89] Davood Rafiei, Krishna Bharat, and Anand Shukla. 2010. Diversifying Web Search Results. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 781–790. https://doi.org/10.1145/1772690.1772770

[90] Karthik Raman, Paul N. Bennett, and Kevyn Collins-Thompson. 2013. Toward Whole-session Relevance: Exploring Intrinsic Diversity in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 463–472. https://doi.org/10.1145/2484028.2484089

[91] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2015. Mining, Ranking and Recommending Entity Aspects. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 263–272. https://doi.org/10.1145/2766462.2767724

[92] Pengjie Ren, Zhumin Chen, Jun Ma, Shuaiqiang Wang, Zhiwei Zhang, Zhaochun Ren, and Tinghuai Ma. 2018. User Session Level Diverse Reranking of Search Results. *Neurocomput.* 274, C (Jan. 2018), 66–79. https://doi.org/10.1016/j.neucom.2016.05.087

[93] Eun Youp Rha, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2016. Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–9. https://doi.org/10.1002/pra2.2016.14505301048

[94] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971), 313–323.

[95] Tuukka Ruotsalo, Kumaripaba Athukorala, Dorota Głowacka, Ksenia Konyushkova, Antti Oulasvirta, Samuli Kaipiainen, Samuel Kaski, and Giulio Jacucci. 2013. Supporting Exploratory Search Tasks with Interactive User Modeling. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries (ASIST '13)*. American Society for Information Science, Silver Springs, MD, USA, Article 39, 10 pages. http://dl.acm.org/citation.cfm?id=2655780.2655819

[96] Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2015. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* 58, 1 (Jan. 2015), 86–92. https://doi.org/10.1145/2656334

[97] Tuukka Ruotsalo, Jaakko Peltonen, Manuel Eugster, Dorota Glowacka, Ksenia Konyushkova, Kumaripaba Athukorala, Ilkka Kosunen Aki Reijonen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2013. Directing Exploratory Search with Interactive Intent Modeling. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 1759–1764. https://doi.org/10.1145/2505515.2505644

[98] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J.A. Eugster, Dorota Głowacka, Aki Reijonen, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2015. SciNet: Interactive Intent Modeling for Information Discovery. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 1043–1044. https://doi.org/10.1145/2766462.2767863

[99] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. 2017. Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 241–250. https://doi.org/10.1109/TVCG.2016.2598495

[100] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. 2010. Clustering Query Refinements by User Intent. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 841–850. https://doi.org/10.1145/1772690.1772776

[101] Gerard Salton and Chris Buckley. 1997. Improving retrieval performance by relevance feedback. *Readings in information retrieval* 24, 5 (1997), 355–363.

[102] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2012. On the role of novelty for search result diversification. *Information Retrieval* 15, 5 (01 Oct 2012), 478–502. https://doi.org/10.1007/s10791-011-9180-x

[103] Walter Schneider and Richard M Shiffrin. 1977. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological review* 84, 1 (1977), 1.

[104] Monica C. Schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. 2006. mSpace: Improving Information Access to Multimedia Domains with Multimodal Exploratory Search. *Commun. ACM* 49, 4 (April 2006), 47–49. https://doi.org/10.1145/1121949.1121980

[105] Marc M. Sebrechts, John V. Cugini, Sharon J. Laskowski, Joanna Vasilakis, and Michael S. Miller. 1999. Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/312624.312634

[106] Milad Shokouhi, Ryen W. White, Paul Bennett, and Filip Radlinski. 2013. Fighting Search Engine Amnesia: Reranking Repeated Results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, ACM, New York, NY, USA, 273–282. https://doi.org/10.1145/2484028.2484075

[107] Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. 2012. Ordinary Search Engine Users Assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)*. ACM, New York, NY, USA, 110–119. https://doi.org/10.1145/2362724.2362746

[108] Mark D. Smucker, Xiaoyu Sunny Guo, and Andrew Toulis. 2014. Mouse Movement During Relevance Judging: Implications for Determining User Attention. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 979–982. https://doi.org/10.1145/2600428.2609489

[109] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 553–562. https://doi.org/10.1145/2806416.2806493

[110] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[111] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

[112] Rohail Syed and Kevyn Collins-Thompson. 2017. Optimizing search results for human learning goals. *Information Retrieval Journal* (12 May 2017). https://doi.org/10.1007/s10791-017-9303-0

[113] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 555–564. https://doi.org/10.1145/3077136.3080835

[114] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The Perfect Search Engine is not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 415–422. https://doi.org/10.1145/985692.985745

[115] Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, and Susan Dumais. 2014. Slow Search. *Commun. ACM* 57, 8 (Aug. 2014), 36–38. https://doi.org/10.1145/2633041

[116] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. 2008. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and development in information retrieval*. ACM, New York, NY, USA, 163–170.

[117] Loren Terveen, Will Hill, and Brian Amento. 1999. Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources. *ACM Transactions on Computer-Human Interaction* 6, 1 (1999), 67–94. https://doi.org/10.1145/310641.310644

[118] Pertti Vakkari. 2003. Task-Based Information Searching. *Annual Review of Information Science and Technology* 37, 1 (2003), 413–464. https://doi.org/10.1002/aris.1440370110

[119] Pertti Vakkari and Saila Huuskonen. 2012. Search Effort Degrades Search Output but Improves Task Outcome. *Journal of the American Society for Information Science and Technology* 63, 4 (2012), 657–670. https://doi.org/10.1002/asi.21683

[120] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. 2010. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *Journal of Machine Learning Research* 11 (2010), 451–490.

[121] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. ACM, ACM, New York, NY, USA, 351–362. https://doi.org/10.1145/2449396.2449442

[122] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 351–362. https://doi.org/10.1145/2449396.2449442

[123] Ryen W. White and Resa A. Roth. 2009. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (March 2009), 1–99.

[124] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization & Comp. Graphics (Proc. InfoVis)* 1, 22 (2016), 649 – 658. http://idl.cs.washington.edu/papers/voyager

[125] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti A. Hearst. 2003. Faceted Metadata for Image Search and Browsing. In *Proceedings of the ACM CHI 2003 Human Factors in Computing Systems Conference*. ACM, New York, NY, USA, 401–408. https://doi.org/10.1145/642611.642681

[126] Yisong Yue and Thorsten Joachims. 2009. Interactively Optimizing Information Retrieval Systems As a Dueling Bandits Problem. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, New York, NY, USA, 1201–1208. https://doi.org/10.1145/1553374.1553527

[127] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems* 22, 2 (2004), 179–214. https://doi.org/10.1145/984321.984322

[128] Sicong Zhang, Jiyun Luo, and Hui Yang. 2014. A POMDP Model for Content-free Document Re-ranking. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 1139–1142. https://doi.org/10.1145/2600428.2609529

Received February 2007; revised March 2009; accepted June 2009